NUCLEIC ACIDS AND PROTEINS FROM CENARCHAEUMSYMBIOSUM

Related Applications

The present application is a divisional of co-pending U.S. Patent Application Serial Number 09/408,020, filed September 29, 1999, which claims priority from U.S. Provisional Patent Application Serial No. 60/102,294, filed September 29, 1998, the disclosure of which is incorporated herein by reference in its entirety.

Background of the Invention

The identification and characterization of organisms which inhabit a diverse range of ecosystems leads to a greater understanding of the operation of such ecosystems. In addition, because the physiology of such organisms is adapted to function in the particular habitat which the organism inhabits, the enzymes which carry out the organism's physiological processes may possess characteristics which provide advantages when they are utilized in therapeutic procedures, industrial applications, or research applications. Furthermore, by determining the sequences of these organisms' genes, insight into their biochemical pathways and processes may be gained without the necessity of culturing the organisms in the laboratory, thereby enabling the physiological characterization of organisms which are recalcitrant to growth in the laboratory. Molecular phylogenetic surveys have recently revealed an ecologically widespread Crenarchaeal group that inhabits cold and temperate terrestrial and marine environments. To date these organisms have resisted isolation in pure culture, so their phenotypic and genotypic characteristics remain largely unknown. In order to characterize the physiology of these archaea, to develop methodological approaches for characterizing uncultivated microorganisms and identifying their presence in a sample, and to identify enzymes produced by these archae which may be useful in therapeutic, industrial, or laboratory applications, genomic analyses of the non-thermophilic crenarchaeote Cenarchaeum symbiosum was undertaken.

Non-thermophilic Crenarchaeota are one of the more abundant, widespread and frequently recovered prokaryotic groups revealed by molecular phylogenetic approaches. These microorganisms were originally detected in high abundance in temperate ocean waters and polar seas. (DeLong, E. F. 1992. Archaea in coastal marine

environments. *Proc. Natl. Acad. Sci.* **89**, 5685-5689; DeLong, E. F *et al.* 1994. High abundance of Archaea in Antarctic marine picoplankton. *Nature* **371**, 695-697; Fuhrman, J. A., *et al.* Davis. 1992. Novel major archaebacterial group from marine plankton. *Nature* 356, 148-149; Massana, R., *et al.* 1997. Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl. Env. Microb.* **63**, 50-56; McInerney, J.O. *et al.* 1995. Recovery and phylogenetic analysis of novel archaeal rRNA sequences from a deep-sea deposit feeder. *Appl. Env. Microb.* **61**, 1646-1648; Preston, C. M. *et al.* 1996. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc. Natl. Acad. Sci.* USA **93,** 6241-6246) Representatives have now been reported in terrestrial environments and freshwater lake sediments, indicating a widespread distribution. (Bintrim, S.B. *et al.* 1997. Molecular phylogeny of Archaea from soil. *Proc. Natl. Acad. Sci.* USA **94**, 277-282; Jurgens, G. *et al.* 1997. Novel group within the kingdom Crenarchaeota from boreal forest soil. *Appl. Env. Mircob.* **63**, 803-80515, Kudo, Y. *et al.* 1997. Peculiar archaea found in Japanese paddy soils. *Biosc. Biotech. Biochem.* **61**, 917-920; Ueda, *et al.* 1995. Molecular phylogenetic analysis of a soil microbial community. *Eur. J. Soil Sci.* **46**, 415-421; Hershberger, K. L. *et al.* 1996. Wide diversity of Crenarchaeota. *Nature* **384**, 420; MacGregor, B.J. 1997. Crenarchaeota in Lake Michigan sediment. *Appl. Env. Microb.* **63**, 1178-1181 *et al.*; Schleper, C.*et al.* 1997. Recovery of crenarchaeotal ribosomal DNA sequences from freshwater-lake sediments. *Appl. Env. Microb.* **63**, 321-323) The ecological distribution of these organisms was initially surprising, since their closest cultivated relatives are all thermophilic or hyperthermophilic. No representative of this new archaeal group has yet been obtained in pure culture, so the phenotypic and metabolic properties of these organisms, as well as their impact on the environment and global nutrient cycling, remain unknown. Since growth temperature and habitat characteristics vary so widely between non-thermophilic and the hyperthermophilic *Creanarchaeota,* these groups are likely to differ greatly with respect to their specific physiology and metabolism.

To gain a better perspective on the genetic and physiological characteristics of non-thermophilic crenarchaeotes, a genomic study of *Cenarchaeum symbiosum* was

begun. This archaeon lives in specific association with the marine sponge *Axinella mexicana* off the coast of California, allowing access to relatively large amounts of biomass from this species. (Preston, C. M. *et al.* 1996. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc. Natl. Acad. Sci.* USA **93**, 6241-6246) The approach taken herein differs in several respects from now standard genomic characterization of cultivated organisms, and also from comparable studies of uncultivated obligate parasites or symbionts. *C. symbiosum* has not been completely physically separated from the tissues of its metazoan host. Therefore, its genetic material needs to be identified within the context of complex genomic libraries that contain significant amounts of eucaryotic DNA, as well as DNA derived from members of *Bacteria*.

Molecular phylogenetic surveys of mixed microbial populations have revealed the existence of many new lineages undetected by classical microbiological approaches. (DeLong, E. F. 1997. Marine microbial diversity: the tip of the iceberg. *Tibtech* **15**, 2-9.; Pace, N. R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734-740) Furthermore, quantitative rRNA hybridization experiments demonstrate that some of these novel prokaryotic groups represent major components of natural microbial communities. These molecular phylogenetic approaches have altered current views of microbial diversity and ecology, and have demonstrated that traditional cultivation techniques may recover only a small, skewed fraction of naturally occurring microbes. However, phylogenetic identification using single gene sequences provides a limited perspective on other biological properties, particularly for novel lineages only distantly related to cultivated and characterized organisms. Consequently, additional approaches are necessary to better characterize ecologically abundant and potentially biotechnologically useful microorganisms, many of which resist cultivation attempts.

Summary of the Invention

One embodiment of the present invention is an isolated, purified, or enriched nucleic acid comprising a sequence selected from the group consisting of SEQ ID NO: 1 and SEQ ID NO: 2, the sequences complementary to SEQ ID NO: 1 and SEQ ID NO: 2, fragments comprising at least 10 consecutive nucleotides of SEQ ID NO: 1 and SEQ ID NO: 2, and fragments comprising at least 10 consecutive nucleotides of the sequences complementary to SEQ ID NO: 1 and SEQ ID NO: 2. One aspect of the

present invention is an isolated, purified, or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of high stringency. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of moderate stringency. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of low stringency. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid having at least 70% homology to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid having at least 99% homology to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters.

Another embodiment of the present invention is an isolated, purified, or enriched nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs: 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79 and the sequences complementary thereto. One aspect of the present invention is an isolated, purified, or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of high stringency. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of moderate stringency. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of low stringency. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid having at least 70% homology to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid having at least 99% homology to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters.

Another embodiment of the present invention is an isolated, purified, or enriched nucleic acid comprising at least 10 consecutive bases of a sequence selected from the group consisting of SEQ ID NOs: 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79 and the sequences complementary thereto. One aspect of

the present invention is an isolated, purified, or enriched nucleic acid having at least 70% homology to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters.

Another embodiment of the present invention is an isolated, purified, or enriched nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs: 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73, 77 and the sequences complementary thereto. One aspect of the present invention is an isolated, purified, or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of high stringency. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of moderate stringency. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid capable of hybridizing to the nucleic acid of this embodiment under conditions of low stringency. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid having at least 70% homology to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid having at least 99% homology to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters.

Another embodiment of the present invention is an isolated, purified, or enriched nucleic acid comprising at least 10 consecutive bases of a sequence selected from the group consisting of SEQ ID NOs: 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73, 77 and the sequences complementary thereto. One aspect of the present invention is an isolated, purified, or enriched nucleic acid having at least 70% homology to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters. Another aspect of the present invention is an isolated, purified, or enriched nucleic acid having at least 99% homology to the nucleic acid of this embodiment as determined by analysis with BLASTN version 2.0 with the default parameters.

Another embodiment of the present invention is an isolated, purified, or enriched nucleic acid encoding a polypeptide having a sequence selected from the group

consisting of SEQ ID NOs: 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, and 80.

Another embodiment of the present invention is an isolated, purified, or enriched nucleic acid encoding a polypeptide comprising at least 10 consecutive amino acids of a polypeptide having a sequence selected from the group consisting of SEQ ID NOs: 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, and 80.

Another embodiment of the present invention is an isolated, purified, or enriched nucleic acid encoding a polypeptide having a sequence selected from the group consisting of SEQ ID NOs: 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

Another embodiment of the present invention is an isolated, purified, or enriched nucleic acid encoding a polypeptide comprising at least 10 consecutive amino acids of a polypeptide having a sequence selected from the group consisting of SEQ ID NOs: 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

Another embodiment of the present invention is an isolated or purified polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, and 80. Another aspect of the present invention is an isolated or purified polypeptide comprising at least 10 consecutive amino acids of the polypeptides of this embodiment. Another aspect of the present invention is an isolated or purified polypeptide having at least 70% homology to the polypeptide of this embodiment as determined by analysis with FASTA version 3.0t78 with the default parameters. Another aspect of the present invention is an isolated or purified polypeptide having at least 99% homology to the polypeptide of this emobdiment as determined by analysis with FASTA version 3.0t78 with the default parameters. Another aspect of the present invention is an isolated or purified polypeptide having at least 70% homology to an isolated or purified polypeptide comprising at least 10 consecutive amino acids of the polypeptides of this embodiment as determined by analysis with FASTA version 3.0t78 with the default parameters. Another aspect of the present invention is an isolated or purified polypeptide having at least 99% homology to the polypeptide of to an isolated or purified polypeptide comprising at least 10 consecutive amino acids of the polypeptides

of this embodiment as determined by analysis with FASTA version 3.0t78 with the default parameters.

Another aspect of the present invention is an isolated or purified polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. One aspect of the present invention is an isolated or purified polypeptide comprising at least 10 consecutive amino acids of the polypeptides of this embodiment. Another aspect of the present invention is an isolated or purified polypeptide having at least 70% homology to the polypeptides of this embodiment as determined by analysis with FASTA version 3.0t78 with the default parameters. Another aspect of the present invention is an isolated or purified polypeptide having at least 99% homology to the polypeptides of this embodiment as determined by analysis with FASTA version 3.0t78 with the default parameters. Another aspect of the present invention is An isolated or purified polypeptide having at least 70% homology to an isolated or purified polypeptide comprising at least 10 consecutive amino acids of the polypeptides of this embodiment as determined by analysis with FASTA version 3.0t78 with the default parameters. Another aspect of the present invention is an isolated or purified polypeptide having at least 99% homology to an isolated or purified polypeptide comprising at least 10 consecutive amino acids of the polypeptides of this embodiment as determined by analysis with FASTA version 3.0t78 with the default parameters.

Another embodiment of the present invention is an isolated or purified antibody capable of specifically binding to a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, and 80.

Another embodiment of the present invention is an isolated or purified antibody capable of specifically binding to a polypeptide comprising at least 10 consecutive amino acids of one of the polypeptides of SEQ ID NOs: 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, and 80.

Another embodiment of the present invention is an isolated or purified antibody capable of specifically binding to a polypeptide having a sequence selected from the group consisting of SEQ ID NOs: 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

Another embodiment of the present invention is an isolated or purified antibody capable of specifically binding to a polypeptide comprising at least 10 consecutive amino acids of one of the polypeptides of SEQ ID NOs: 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

Another embodiment of the present invention is a method of making a polypeptide having a sequence selected from the group consisting of SEQ ID NOs: 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, and 80 comprising introducing a nucleic acid encoding said polypeptide, said nucleic acid being operably linked to a promoter, into a host cell.

Another embodiment of the present invention is a method of making a polypeptide comprising at least 10 amino acids of a sequence selected from the group consisting of the sequences of SEQ ID NOs: 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, and 80 comprising introducing a nucleic acid encoding said polypeptide, said nucleic acid being operably linked to a promoter, into a host cell.

Another embodiment of the present invention is a method of making a polypeptide having a sequence selected from the group consisting of SEQ ID NOs: 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 comprising introducing a nucleic acid encoding said polypeptide, said nucleic acid being operably linked to a promoter, into a host cell.

Another embodiment of the present invention is a method of making a polypeptide comprising at least 10 amino acids of a sequence selected from the group consisting of the sequences of SEQ ID NOs: 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 comprising introducing a nucleic acid encoding said polypeptide, said nucleic acid being operably linked to a promoter, into a host cell.

Another embodiment of the present i method of generating a variant comprising obtaining a nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77, the sequences complementary to the sequences of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 , fragments comprising at least 30 consecutive nucleotides of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63,

65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77, and fragments comprising at least 30 consecutive nucleotides of the sequences complementary to SEQ ID NOS. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 and changing one or more nucleotides in said sequence to another nucleotide, deleting one or more nucleotides in said sequence, or adding one or more nucleotides to said sequence. In one aspect of the present invention, the method further comprises the step of testing the enzymatic properties of a translation product of said variant.

Another embodiment of the present invention is a computer readable medium having stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 and a polypeptide code of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

Another embodiment of the present invention is a computer system comprising a processor and a data storage device wherein said data storage device has stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 and a polypeptide code of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. In one aspect of the present invention, the computer system further comprises a sequence comparer and a data storage device having reference sequences stored thereon. For example, the sequence comparer may comprise a computer program which indicates polymorphisms. In another aspect of the present invention is the computer system of this embodiment further comprises an identifier which identifies features in said sequence.

Another embodiment of the present invention is a method for comparing a first sequence to a reference sequence wherein said first sequence is selected from the group consisting of a nucleic acid code of SEQID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 and a polypeptide code of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32,

34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 comprising the steps of reading said first sequence and said reference sequence through use of a computer program which compares sequences; and determining differences between said first sequence and said reference sequence with said computer program. In one aspect of the present invention, the step of determining differences between the first sequence and the reference sequence comprises identifying polymorphisms.

Another embodiment of the present invention is a method for identifying a feature in a sequence selected from the group consisting of a nucleic acid code of SEQID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 and a polypeptide code of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 comprising the steps of reading said sequence through the use of a computer program which identifies features in sequences and identifying features in said sequence with said computer program.

Brief Description of the Drawings

Figure 1 shows the locations of coding regions, the %G-C. and the %DNA identity between the approximately 28Kb of common sequence in fosmids 101G10 and 60A5.

Figure 2 shows the sequences surrounding the TATA boxes of several promoters from *Cenarchaeum symbiosum* and the distances from the TATA boxes to the initiation codons in these sequences.

Figure 3 is a block diagram of an exemplary computer system.

Figure 4 is a flow diagram illustrating one embodiment of a process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the homology levels between the new sequence and the sequences in the database.

Figure 5 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous.

Figure 6 is a flow diagram illustrating one embodiment of an identifier process for detecting the presence of a feature in a sequence.

Definitions

The term "gene" means the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as, where applicable, intervening sequences (introns) between individual coding segments (exons).

As used herein, the term "isolated" means that the material is removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotides could be part of a vector and/or such polynucleotides or polypeptides could be part of a composition, and still be isolated in that such vector or composition is not part of its natural environment.

As used herein, the term "purified" does not require absolute purity; rather, it is intended as a relative definition. Individual nucleic acids obtained from a library have been conventionally purified to electrophoretic homogeneity. The sequences obtained from these clones could not be obtained directly either from the library or from total human DNA. The purified nucleic acids of the present invention have been purified from the remainder of the genomic DNA in the organism by at least 10^4-10^6 fold. However, the term "purified" also includes nucleic acids which have been purified from the remainder of the genomic DNA or from other sequences in a library or other environment by at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude.

As used herein, the term "recombinant" means that the nucleic acid is adjacent to "backbone" nucleic acid to which it is not adjacent in its natural environment. Additionally, to be "enriched" the nucleic acids will represent 5% or more of the number of nucleic acid inserts in a population of nucleic acid backbone molecules. Backbone molecules according to the present invention include nucleic acids such as expression vectors, self-replicating nucleic acids, viruses, integrating nucleic acids, and other vectors or nucleic acids used to maintain or manipulate a nucleic acid insert of interest. Preferably,

the enriched nucleic acids represent 15% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. More preferably, the enriched nucleic acids represent 50% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. In a highly preferred embodiment, the enriched nucleic acids represent 90% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules.

A promoter sequence is "operably linked to" a coding sequence when RNA polymerase which initiates transcription at the promoter will transcribe the coding sequence into mRNA.

"Recombinant" polypeptides or proteins refer to polypeptides or proteins produced by recombinant DNA techniques; *i.e.*, produced from cells transformed by an exogenous DNA construct encoding the desired polypeptide or protein. "Synthetic" polypeptides or protein are those prepared by chemical synthesis.

A DNA "coding sequence" or a "nucleotide sequence encoding" a particular polypeptide or protein, is a DNA sequence which is transcribed and translated into a polypeptide or protein when placed under the control of appropriate regulatory sequences.

"Plasmids" are designated by a lower case p preceded and/or followed by capital letters and/or numbers. The starting plasmids herein are either commercially available, publicly available on an unrestricted basis, or can be constructed from available plasmids in accord with published procedures. In addition, equivalent plasmids to those described herein are known in the art and will be apparent to the ordinarily skilled artisan.

"Digestion" of DNA refers to catalytic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in the DNA. The various restriction enzymes used herein are commercially available and their reaction conditions, cofactors and other requirements were used as would be known to the ordinarily skilled artisan. For analytical purposes, typically 1 μg of plasmid or DNA fragment is used with about 2 units of enzyme in about 20 μl of buffer solution. For the purpose of isolating DNA fragments for plasmid construction, typically 5 to 50 μg of DNA are digested with 20 to 250 units of enzyme in a larger volume. Appropriate buffers and substrate amounts for particular restriction enzymes are specified by the manufacturer. Incubation times of

about 1 hour at 37°C are ordinarily used, but may vary in accordance with the supplier's instructions. After digestion the gel electrophoresis may be performed to isolate the desired fragment.

"Oligonucleotide" refers to either a single stranded polydeoxynucleotide or two complementary polydeoxynucleotide strands which may be chemically synthesized. Such synthetic oligonucleotides have no 5' phosphate and thus will not ligate to another oligonucleotide without adding a phosphate with an ATP in the presence of a kinase. A synthetic oligonucleotide will ligate to a fragment that has not been dephosphorylated.

Detailed Description of the Preferred Embodiment

In order to begin the characterization of *Cenarchaeum symbiosum*, a large region of the *C. symbiosum* genome was sequenced. In particular, two overlapping *C. symbiosum*-derived fosmid inserts of approximately 42kb and 33kb were sequenced. The sequences of the two fosmid inserts revealed that there are at least two major variants or strains of *C. symbiosum* that coexist inside the sponge tissues of a single sponge. This complexity of the *C. symbiosum* population was not detected in initial studies based solely on direct sequencing of PCR amplified SSU genes. (Preston, C. M. *et al.* 1996. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc. Natl. Acad. Sci.* USA **93,** 6241-6246) This natural variation would also have been lost upon isolation of a pure culture.

The *Cenarchaeum symbiosum* sequences obtained from the two fosmids containing overlapping genomic inserts are provided in the accompanying sequence listing and are identified as SEQ ID NO: 1 and SEQ ID NO: 2. The two fosmid sequences were not entirely identical in their overlapping portions but instead contained differences. Upon further investigation, it was discovered that the two fosmid sequences were derived from two different, but closely related, strains of *Cenarchaeum symbiosum* (called variant A and variant B) which may simultaneously inhabit a single sponge.

Within the sequences of the fosmid inserts, numerous open reading frames encoding polypeptides having homology to known proteins, as well as open reading frames encoding proteins which do not exhibit homology to known proteins, were identified. Homology was determined using the program FASTA with the default

parameters. The polypeptides encoded by these sequences are identified in the accompanying sequence listing as SEQ ID NOs: 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76 and 80 (polypeptides with homology to known proteins) and SEQ ID NOs: 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74 and 78 (polypeptides without homology to known proteins). In addition, sequences encoding the 16S rRNA, the 23S rRNA and a tyrosine tRNAs were also identified.

One aspect of the present invention is an isolated, purified, or enriched nucleic acid comprising one of the sequences of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 the sequences complementary thereto, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive bases of one of the sequences of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 or the sequences complementary thereto. The isolated, purified or enriched nucleic acids may comprise DNA, including cDNA, genomic DNA, and synthetic DNA. The DNA may be double-stranded or single-stranded, and if single stranded may be the coding strand or non-coding (anti-sense) strand. Alternatively, the isolated, purified or enriched nucleic acids may comprise RNA.

As discussed in more detail below, the isolated, purified, or enriched nucleic acids of one of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 may be used to prepare one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80.

Accordingly, another aspect of the present invention is an isolated, purified, or enriched nucleic acid which encodes one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56,

58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80. The coding sequences of these nucleic acids may be identical to one of the coding sequences of one of the nucleic acids of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 or a fragment thereof or may be different coding sequences which encode one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 as a result of the redundancy or degeneracy of the genetic code. The genetic code is well known to those of skill in the art and can be obtained, for example, on page 214 of B. Lewin, Genes VI, Oxford University Press, 1997, the disclosure of which is incorporated herein by reference.

The isolated, purified, or enriched nucleic acid which encodes one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 may include, but is not limited to: only the coding sequence of one of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79; the coding sequences of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 and additional coding sequences, such as leader sequences or proprotein sequences; or the coding sequences of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 and non-coding sequences, such as introns or non-coding sequences 5' and/or 3' of the coding sequence. Thus, as used herein, the term "polynucleotide encoding a polypeptide" encompasses a polynucleotide which includes only coding sequence for

the polypeptide as well as a polynucleotide which includes additional coding and/or non-coding sequence.

Alternatively, the nucleic acid sequences of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 may be mutagenized using conventional techniques, such as site directed mutagenesis, or other techniques familiar to those skilled in the art, to introduce silent changes into the polynucleotides of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79. As used herein, "silent changes" include, for example, changes which do not alter the amino acid sequence encoded by the polynucleotide. Such changes may be desirable in order to increase the level of the polypeptide produced by host cells containing a vector encoding the polypeptide by introducing codons or codon pairs which occur frequently in the host organism.

The present invention also relates to polynucleotides which have nucleotide changes which result in amino acid substitutions, additions, deletions, fusions and truncations in the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80. Such nucleotide changes may be introduced using techniques such as site directed mutagenesis, random chemical mutagenesis, exonuclease III deletion, and other recombinant DNA techniques. Alternatively, such nucleotide changes may be naturally occurring allelic variants which are isolated by identifying nucleic acids which specifically hybridize to probes comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive bases of one of the sequences of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 or the sequences complementary thereto to nucleic acids from *Cenarchaeum symbiosum* or related organisms under conditions of high, moderate, or low strigency as provided herein.

The isolated, purified, or enriched nucleic acids of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79, the sequences complementary thereto, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300,

400, or 500 consecutive bases of one of the sequences of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 or the sequences complementary thereto may also be used as probes to identify the presence of *Cenarchaeum symbiosum* in a biological sample. In such procedures, a biological sample potentially harboring *Cenarchaeum symbiosum* is obtained and nucleic acids are obtained from the sample. The nucleic acids are contacted with the probe under conditions which permit the probe to specifically hybridize to any complementary sequences from *Cenarchaeum symbiosum* which are present therein.

Where necessary, conditions which permit the probe to specifically hybridize to complementary sequences from *Cenarchaeum symbiosum* may be determined by placing the probe in contact with complementary sequences from *Cenarchaeum symbiosum* as well as control sequences which are not from *Cenarchaeum symbiosum.* In some analyses, the control sequences may be from organisms related to *Cenarchaeum symbiosum.* Alternatively, the control sequences may be from organisms which are not related to *Cenarchaeum symbiosum.* Hybridization conditions, such as the salt concentration of the hybridization buffer, the formamide concentration of the hybridization buffer, or the hybridization temperature, may be varied to identify conditions which allow the probe to hybridize specifically to nucleic acids from *Cenarchaeum symbiosum.*

If the sample contains nucleic acids from *Cenarchaeum symbiosum,* specific hybridization of the probe to the nucleic acids from *Cenarchaeum symbiosum* is then detected. Hybridization may be detected by labeling the probe with a detectable agent such as a radioactive isotope, a fluorescent dye or an enzyme capable of catalyzing the formation of a detectable product.

Many methods for using the labeled probes to detect the presence of nucleic acids from *Cenarchaeum symbiosum* in a sample are familiar to those skilled in the art. These include Southern Blots, Northern Blots, colony hybridization procedures, and dot blots. Protocols for each of these procedures are provided in Ausubel et al. Current Protocols in Molecular Biology, John Wiley 503 Sons, Inc. 1997 and Sambrook et al., Molecular Cloning: A Laboratory Manual 2d Ed., Cold Spring Harbor Laboratory Press, 1989, the entire disclosures of which are incorporated herein by reference.

Alternatively, more than one probe (at least one of which is capable of specifically hybridizing to any complementary sequences from *Cenarchaeum symbiosum* which are present in the nucleic acid sample), may be used in an amplification reaction to determine whether the nucleic acid sample contains nucleic acids from *Cenarchaeum symbiosum*. Preferably, the probes comprise oligonucleotides. In one embodiment, the amplification reaction may comprise a PCR reaction. PCR protocols are described in Ausubel and Sambrook, *supra*. Alternatively, the amplification may comprise a ligase chain reaction, 3SR, or strand displacement reaction. (See Barany, F., "The Ligase Chain Reaction in a PCR World", *PCR Methods and Applications* 1:5-16 (1991); E. Fahy *et al.*, "Self-sustained Sequence Replication (3SR): An Isothermal Transcription-based Amplification System Alternative to PCR", *PCR Methods and Applications* 1:25-33 (1991); and Walker G.T. *et al.*, "Strand Displacement Amplification-an Isothermal *in vitro* DNA Amplification Technique, *Nucleic Acid Research* 20:1691-1696 (1992) the disclosures of which are incorporated herein by reference in their entireties). In such procedures, the nucleic acids in the sample are contacted with the probes, the amplification reaction is performed, and any resulting amplification product is detected. The amplification product may be detected by performing gel electrophoresis on the reaction products and staining the gel with an interculator such as ethidium bromide. Alternatively, one or more of the probes may be labeled with a radioactive isotope and the presence of a radioactive amplification product may be detected by autoradiography after gel electrophoresis.

Probes derived from sequences near the ends of the sequences of SEQ ID Nos: 1 and 2 may also be used in chromosome walking procedures to identify clones containing genomic sequences located adjacent to the sequences of SEQ ID Nos: 1 and 2. Such methods allow the isolation of genes which encode additional proteins expressed in *Cenarchaeum symbiosum* and facilitate the further physiological characterization of the organism.

Another aspect of the present invention is a method for determining whether a sample contains variant A and/or variant B of *Cenarchaeum symbiosum*. In such procedures, a sample potentially harboring variant A and/or variant B *Cenarchaeum symbiosum* is obtained and nucleic acids are obtained from the sample. The nucleic acids are contacted with the probe under conditions which permit the probe to

specifically hybridize to any complementary sequences from variant A or variant B of *Cenarchaeum symbiosum* which are present therein. Preferably, the probe comprises a sequence having one or more nucleotides which differ between variant A and variant B. Conditions in which the probe specifically hybridizes to nucleic acids from one of the variants but not to nucleic acids from the other variant may be determined by contacting the probe with its corresponding sequence from variant A and variant B and varying the hybridization conditions, such as the salt concentration of the hybridization buffer, the formamide concentration of the buffer, or the hybridization temperature, to identify conditions in which the probe hybridizes to the corresponding sequence from one variant but not to the corresponding sequence from the other variant. Hybridization of the probe to nucleic acids from the *Cenarchaeum symbiosum* variant is then detected using any of the procedures described above.

The isolated, purified, or enriched nucleic acids of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79, the sequences complementary thereto, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive bases of one of the sequences of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 or the sequences complementary thereto may be used as probes to identify and isolate cDNAs encoding the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80. In such procedures, a cDNA library is constructed from a sample containing *Cenarchaeum symbiosum*. The cDNA library is then contacted with a probe comprising a coding sequence, or a fragment of a coding sequence, encoding one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or a fragment thereof under conditions which permit the probe to specifically hybridize to sequences complementary thereto. cDNAs which hybridize to the probe are then detected and isolated. Procedures for preparing and identifying cDNAs are disclosed in Ausubel et al. Current Protocols in Molecular Biology, John Wiley 503 Sons, Inc. 1997 and Sambrook et al., Molecular Cloning: A Laboratory Manual 2d Ed., Cold Spring Harbor Laboratory Press, 1989, the

disclosures of which are incorporated herein by reference.

The isolated, purified, or enriched nucleic acids of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79, the sequences complementary thereto, or a fragment comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive bases of one of the sequences of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 or the sequences complementary thereto may be used as probes to identify and isolate related nucleic acids. In some embodiments, the related nucleic acids may be cDNAs or genomic DNAs from organisms other than *Cenarchaeum symbiosum*. For example, the other organisms may be organisms which are related to *Cenarchaeum symbiosum*. In such procedures, a nucleic acid sample containing nucleic acids from the related organism, such as a cDNA or genomic DNA library from the related organism, is contacted with the probe under conditions which permit the probe to specifically hybridize to related sequences. Hybridization of the probe to nucleic acids from the related organism is then detected using any of the methods described above.

Hybridization may be carried out under conditions of low stringency, moderate stringency or high stringency. As an example of nucleic acid hybridization, a polymer membrane containing immobilized denatured nucleic acids is first prehybridized for 30 minutes at 45°C in a solution consisting of 0.9 M NaCl, 50 mM NaH_2PO_4, pH 7.0, 5.0 mM Na_2EDTA, 0.5% SDS, 10X Denhardt's, and 0.5 mg/ml polyriboadenylic acid. Approximately 2 X 10^7 cpm (specific activity 4-9 X 10^8 cpm/ug) of ^{32}P end-labeled oligonucleotide probe are then added to the solution. After 12-16 hours of incubation, the membrane is washed for 30 minutes at room temperature in 1X SET (150 mM NaCl, 20 mM Tris hydrochloride, pH 7.8, 1 mM Na_2EDTA) containing 0.5% SDS, followed by a 30 minute wash in fresh 1X SET at Tm-10°C for the oligonucleotide probe. The membrane is then exposed to auto-radiographic film for detection of hybridization signals.

By varying the stringency of the hybridization conditions used to identify nucleic acids, such as cDNAs or genomic DNAs, which hybridize to the detectable probe, nucleic acids having different levels of homology to the probe can be identified and isolated.

Stringency may be varied by conducting the hybridization at varying temperatures below the melting temperatures of the probes. The melting temperature of the probe may be calculated using the following formulas:

For probes between 14 and 70 nucleotides in length the melting temperature (Tm) is calculated using the formula: $Tm=81.5+16.6(\log [Na+])+0.41(\text{fraction G+C})-(600/N)$ where N is the length of the probe.

If the hybridization is carried out in a solution containing formamide, the melting temperature may be calculated using the equation $Tm=81.5+16.6(\log [Na+])+0.41(\text{fraction G+C})-(0.63\% \text{ formamide})-(600/N)$ where N is the length of the probe.

Prehybridization may be carried out in 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 100μg denatured fragmented salmon sperm DNA or 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 100μg denatured fragmented salmon sperm DNA, 50% formamide. The formulas for SSC and Denhardt's solutions are listed in Sambrook et al., *supra.*

Hybridization is conducted by adding the detectable probe to the prehybridization solutions listed above. Where the probe comprises double stranded DNA, it is denatured before addition to the hybridization solution. The filter is contacted with the hybridization solution for a sufficient period of time to allow the probe to hybridize to cDNAs or genomic DNAs containing sequences complementary thereto or homologous thereto. For probes over 200 nucleotides in length, the hybridization may be carried out at 15-25°C below the Tm. For shorter probes, such as oligonucleotide probes, the hybridization may be conducted at 5-10°C below the Tm. Preferably, for hybridizations in 6X SSC, the hybridization is conducted at approximately 68°C. Preferably, for hybridizations in 50% formamide containing solutions, the hybridization is conducted at approximately 42°C.

All of the foregoing hybridizations would be considered to be under conditions of high stringency.

Following hybridization, the filter is washed in 2X SSC, 0.1% SDS at room temperature for 15 minutes. The filter is then washed with 0.1X SSC, 0.5% SDS at room temperature for 30 minutes to 1 hour. Thereafter, the solution is washed at the hybridization temperature in 0.1X SSC, 0.5% SDS. A final wash is conducted in 0.1X SSC at room temperature.

Nucleic acids which have hybridized to the probe are identified by autoradiography or other conventional techniques.

The above procedure may be modified to identify nucleic acids having decreasing levels of homology to the probe sequence. For example, to obtain nucleic acids of decreasing homology to the detectable probe, less stringent conditions may be used. For example, the hybridization temperature may be decreased in increments of 5°C from 68°C to 42°C in a hybridization buffer having a Na+ concentration of approximately 1M. Following hybridization, the filter may be washed with 2X SSC, 0.5% SDS at the temperature of hybridization. These conditions are considered to be "moderate" conditions above 50°C and "low" conditions below 50°C. A specific example of "moderate" hybridization conditions is when the above hybridization is conducted at 55°C. A specific example of "low stringency" hybridization conditions is when the above hybridization is conducted at 45°C.

Alternatively, the hybridization may be carried out in buffers, such as 6X SSC, containing formamide at a temperature of 42°C. In this case, the concentration of formamide in the hybridization buffer may be reduced in 5% increments from 50% to 0% to identify clones having decreasing levels of homology to the probe. Following hybridization, the filter may be washed with 6X SSC, 0.5% SDS at 50°C. These conditions are considered to be "moderate" conditions above 25% formamide and "low" conditions below 25% formamide. A specific example of "moderate" hybridization conditions is when the above hybridization is conducted at 30% formamide. A specific example of "low stringency" hybridization conditions is when the above hybridization is conducted at 10% formamide.

Nucleic acids which have hybridized to the probe are identified by autoradiography.

For example, the preceding methods may be used to isolate nucleic acids having a sequence with at least 97%, at least 95%, at least 90%, at least 85%, at least 80%, or at least 70% homology to a nucleic acid sequence selected from the group consisting of one of the sequences of SEQ ID NOS. 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79, fragments comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive bases thereof, and the sequences complementary thereto. Homology may be measured using BLASTN version 2.0 with the default parameters. For example, the homologous polynucleotides may have a coding sequence

which is a naturally occurring allelic variant of one of the coding sequences described herein. Such allelic variants may have a substitution, deletion or addition of one or more nucleotides when compared to the nucleic acids of SEQ ID NOs: 1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77 and 79 or the sequences complementary thereto.

Additionally, the above procedures may be used to isolate nucleic acids which encode polypeptides having at least 99%, 95%, at least 90%, at least 85%, at least 80%, or at least 70% homology to a polypeptide having the sequence of one of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof as determined using the FASTA version 3.0t78 algorithm with the default parameters.

Another aspect of the present invention is an isolated or purified polypeptide comprising the sequence of one of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. As discussed above, such polypeptides may be obtained by inserting a nucleic acid encoding the polypeptide into a vector such that the coding sequence is operably linked to a sequence capable of driving the expression of the encoded polypeptide in a suitable host cell. For example, the expression vector may comprise a promoter, a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression.

Promoters suitable for expressing the polypeptide or fragment thereof in bacteria include the E. coli. lac or trp promoters, the lacI promoter, the lacZ promoter, the T3 promoter, the T7 promoter, the gpt promoter, the lambda P_R promoter, the lambda P_L promoter the trp promoter, promoters from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), and the acid phosphatase promoter. Fungal promoters include the α factor promoter. Eukaryotic promoters include the CMV immediate early promoter, the HSV thymidine kinase promoter, heat shock promoters, the early and late SV40 promoter, LTRs from retroviruses, and the mouse

metallothionein-I promoter. Other promoters known to control expression of genes in prokaryotic or eukaryotic cells or their viruses may also be used.

Mammalian expression vectors may also comprise an origin of replication, any necessary ribosome binding sites, a polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. In some embodiments, DNA sequences derived from the SV40 splice and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

Vectors for expressing the polypeptide or fragment thereof in eukaryotic cells may also contain enhancers to increase expression levels. Enhancers are cis-acting elements of DNA, usually from about 10 to about 300 bp in length that act on a promoter to increase its transcription. Examples include the SV40 enhancer on the late side of the replication origin bp 100 to 270, the cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and the adenovirus enhancers.

In addition, the expression vectors preferably contain one or more selectable marker genes to permit selection of host cells containing the vector. Such selectable markers include genes encoding dihydrofolate reductase or genes conferring neomycin resistance for eukaryotic cell culture, genes conferring tetracycline or ampicillin resistance in E. coli, and the S. cerevisiae TRP1 gene.

In some embodiments, the nucleic acid encoding one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof is assembled in appropriate phase with a leader sequence capable of directing secretion of the translated polypeptide or fragment thereof. Optionally, the nucleic acid can encode a fusion polypeptide in which one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof is fused to heterologous peptides or polypeptides, such as N-terminal identification peptides which impart desired characteristics, such as increased stability or simplified purification.

The appropriate DNA sequence may be inserted into the vector by a variety of procedures. In general, the DNA sequence is ligated to the desired position in the vector following digestion of the insert and the vector with appropriate restriction endonucleases. Alternatively, blunt ends in both the insert and the vector may be ligated. A variety of cloning techniques are disclosed in Ausubel et al. Current Protocols in Molecular Biology, John Wiley 503 Sons, Inc. 1997 and Sambrook et al., Molecular Cloning: A Laboratory Manual 2d Ed., Cold Spring Harbor Laboratory Press, 1989, the entire disclosures of which are incorporated herein by reference. Such procedures and others are deemed to be within the scope of those skilled in the art.

The vector may be, for example, in the form of a plasmid, a viral particle, or a phage. Other vectors include chromosomal, nonchromosomal and synthetic DNA sequences, derivatives of SV40; bacterial plasmids, phage DNA, baculovirus, yeast plasmids, vectors derived from combinations of plasmids and phage DNA, viral DNA such as vaccinia, adenovirus, fowl pox virus, and pseudorabies. A variety of cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook, et al., Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor, N.Y., (1989), the disclosure of which is hereby incorporated by reference.

Particular bacterial vectors which may be used include the commercially available plasmids comprising genetic elements of the well known cloning vector pBR322 (ATCC 37017), pKK223-3 (Pharmacia Fine Chemicals, Uppsala, Sweden), GEM1 (Promega Biotec, Madison, WI, USA) pQE70, pQE60, pQE-9 (Qiagen), pD10, psiX174 pBluescript II KS, pNH8A, pNH16a, pNH18A, pNH46A (Stratagene), ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia), pKK232-8 and pCM7. Particular eukaryotic vectors include pSV2CAT, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, and pSVL (Pharmacia). However, any other vector may be used as long as it is replicable and viable in the host cell.

The host cell may be any of the host cells familiar to those skilled in the art, including prokaryotic cells, eukaryotic cells, mammalian cells, insect cells, or plant cells. As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as E. coli, Streptomyces, Bacillus subtilis, Salmonella typhimurium and various species within the genera Pseudomonas, Streptomyces, and Staphylococcus,

fungal cells, such as yeast, insect cells such as Drosophila S2 and Spodoptera Sf9, animal cells such as CHO, COS or Bowes melanoma, and adenoviruses. The selection of an appropriate host is within the abilities of those skilled in the art.

The vector may be introduced into the host cells using any of a variety of techniques, including transformation, transfection, transduction, viral infection, gene guns, or Ti-mediated gene transfer. Particular methods include calcium phosphate transfection, DEAE-Dextran mediated transfection, lipofection, or electroporation (Davis, L., Dibner, M., Battey, I., Basic Methods in Molecular Biology, (1986)).

Where appropriate, the engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying the genes of the present invention. Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter may be induced by appropriate means (e.g., temperature shift or chemical induction) and the cells may be cultured for an additional period to allow them to produce the desired polypeptide or fragment thereof.

Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract is retained for further purification. Microbial cells employed for expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents. Such methods are well known to those skilled in the art. The expressed polypeptide or fragment thereof can be recovered and purified from recombinant cell cultures by methods including ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. Protein refolding steps can be used, as necessary, in completing configuration of the polypeptide. If desired, high performance liquid chromatography (HPLC) can be employed for final purification steps.

Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts (described by Gluzman, Cell, 23:175 (1981), and

other cell lines capable of expressing proteins from a compatible vector, such as the C127, 3T3, CHO, HeLa and BHK cell lines.

The constructs in host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence. Depending upon the host employed in a recombinant production procedure, the polypeptides produced by host cells containing the vector may be glycosylated or may be non-glycosylated. Polypeptides of the invention may or may not also include an initial methionine amino acid residue.

Alternatively, the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof can be synthetically produced by conventional peptide synthesizers. In other embodiments, fragments or portions of the polypeptides may be employed for producing the corresponding full-length polypeptide by peptide synthesis; therefore, the fragments may be employed as intermediates for producing the full-length polypeptides.

Cell-free translation systems can also be employed to produce one of the polypeptides of SEQ ID Nos: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof using mRNAs transcribed from a DNA construct comprising a promoter operably linked to a nucleic acid encoding the polypeptide or fragment thereof. In some embodiments, the DNA construct may be linearized prior to conducting an *in vitro* transcription reaction. The transcribed mRNA is then incubated with an appropriate cell-free translation extract, such as a rabbit reticulocyte extract, to produce the desired polypeptide or fragment thereof.

The present invention also relates to variants of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. The term "variant" includes derivatives or analogs of these polypeptides. In particular, the variants may differ in amino acid sequence from the polypeptides of SEQ ID NOs:

4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 by one or more substitutions, additions, deletions, fusions and truncations, which may be present in any combination.

The variants may be naturally occurring or created in vitro. In particular, such variants may be created using genetic engineering techniques such as site directed mutagenesis, random chemical mutagenesis, Exonuclease III deletion procedures, and standard cloning techniques. Alternatively, such variants, fragments, analogs, or derivatives may be created using chemical synthesis or modification procedures.

Other methods of making variants are also familiar to those skilled in the art. These include procedures in which nucleic acid sequences obtained from natural isolates are modified to generate nucleic acids which encode polypeptides having characteristics which enhance their value in industrial or laboratory applications. In such procedures, a large number of variant sequences having one or more nucleotide differences with respect to the sequence obtained from the natural isolate are generated and characterized. Preferably, these nucleotide differences result in amino acid changes with respect to the polypeptides encoded by the nucleic acids from the natural isolates.

For example, variants may be created using error prone PCR. In error prone PCR, PCR is performed under conditions where the copying fidelity of the DNA polymerase is low, such that a high rate of point mutations is obtained along the entire length of the PCR product. Error prone PCR is described in Leung, D.W., *et al.*, Technique, 1:11-15 (19 89) and Caldwell, R. C. & Joyce G.F., PCR Methods Applic., 2:28-33 (1992), the disclosure of which is incorporated herein by reference in its entirety. Briefly, in such procedures, nucleic acids to be mutagenized are mixed with PCR primers, reaction buffer, $MgCl_2$, $MnCl_2$, Taq polymerase and an appropriate concentration of dNTPs for achieving a high rate of point mutation along the entire length of the PCR product. For example, the reaction may be performed using 20 fmoles of nucleic acid to be mutagenized, 30pmole of each PCR primer, a reaction buffer comprising 50mM KCl, 10mM Tris HCl (pH 8.3) and 0.01% gelatin, 7mM $MgCl_2$, 0.5mM $MnCl_2$, 5 units of Taq polymerase, 0.2mM dGTP, 0.2mM dATP, 1mM dCTP, and 1mM dTTP. PCR may be performed for 30 cycles of 94° C for 1 min, 45° C for 1 min, and 72° C for 1 min. However, it will be appreciated that these parameters may be varied as appropriate. The mutagenized nucleic acids are cloned into an

appropriate vector and the activities of the polypeptides encoded by the mutagenized nucleic acids is evaluated.

Variants may also be created using oligonucleotide directed mutagenesis to generate site-specific mutations in any cloned DNA segment of interest. Oligonucleotide mutagenesis is described in Reidhaar-Olson, J.F. & Sauer, R.T., *et al.*, Science, 241:53-57 (1988), the disclosure of which is incorporated herein by reference in its entirety. Briefly, in such procedures a plurality of double stranded oligonucleotides bearing one or more mutations to be introduced into the cloned DNA are synthesized and inserted into the cloned DNA to be mutagenized. Clones containing the mutagenized DNA are recovered and the activities of the polypeptides they encode are assessed.

Another method for generating variants is assembly PCR. Assembly PCR involves the assembly of a PCR product from a mixture of small DNA fragments. A large number of different PCR reactions occur in parallel in the same vial, with the products of one reaction priming the products of another reaction. Assembly PCR is described in U.S. Patent Application Serial No. 08/677,112, filed July 9, 1997 and U.S. Patent Application Serial No. 08/942,504, filed October 31, 1997, the disclosures of which are incorporated herein by reference in their entireties.

Still another method of genrating variants is sexual PCR mutagenesis. In sexual PCR mutagenesis, forced homologous recombination occurs between DNA molecules of different but highly related DNA sequence in vitro, as a result of random fragmentation of the DNA molecule based on sequence homology, followed by fixation of the crossover by primer extension in a PCR reaction. Sexual PCR mutagenesis is described in Stemmer, W.P., PNAS, USA, 91:10747-10751 (1994), the disclosure of which is incorporated herein by reference. Briefly, in such procedures a plurality of nucleic acids to be recombined are digested with DNAse to generate fragments having an average size of 50-200 nucleotides. Fragments of the desired average size are purifed and resuspended in a PCR mixture. PCR is conducted under conditions which facilitate recombination between the nucleic acid fragments. For example, PCR may be performed by resuspending the purified fragments at a concentration of 10-30ng/μl in a solution of 0.2mM of each dNTP, 2.2mM MgCl2, 50mM KCL, 10mM Tris HCl, pH 9.0, and 0.1% Triton X-100. 2.5 units of Taq polymerase per 100μl of reaction mixture

is added and PCR is performed using the following regime: 94° C for 60 seconds, 94° C for 30 seconds, 50-55° C for 30 seconds, 72° C for 30 seconds (30-45 times) and 72° C for 5 minutes. However, it will be appreciated that these parameters may be varied as appropriate. In some embodiments, oligonucleotides may be included in the PCR reactions. In other embodiments, the Klenow fragment of DNA polymerase I may be used in a first set of PCR reactions and Taq polymerase may be used in a subsequent set of PCR reactions. Recombinant sequences are isolated and the activities of the polypeptides they encode are assessed.

Variants may also be created by in vivo mutagenesis. In some embodiments, random mutations in a sequence of interest are generated by propagating the sequence of interest in a bacterial strain, such as an E. coli strain, which carries mutations in one or more of the DNA repair pathways. Such "mutator" strains have a higher random mutation rate than that of a wild-type parent. Propagating the DNA in one of these strains will eventually generate random mutations within the DNA. Mutator strains suitable for use for in vivo mutagenesis are described in PCT Published Application WO 91/16427, the disclosure of which is incorporated herein by reference in its entirety.

Variants may also be generated using cassette mutagenesis. In cassette mutagenesis a small region of a double stranded DNA molecule is replaced with a synthetic oligonucleotide "cassette" that differs from the native sequence. The oligonucleotide often contains completely and/or partially randomized native sequence.

Recursive ensemble mutagenesis may also be used to generate variants. Recursive ensemble mutagenesis is an algorithm for protein engineering (protein mutagenesis) developed to produce diverse populations of phenotypically related mutants whose members differ in amino acid sequence. This method uses a feedback mechanism to control successive rounds of combinatorial cassette mutagenesis. Recursive ensemble mutagenesis is described in Arkin, A.P. and Youvan, D.C., PNAS, USA, 89:7811-7815 (1992), the disclosure of which is incorporated herein by reference in its entirety.

In some embodiments, variants are created using exponential ensemble mutagenesis. Exponential ensemble mutagenesis is a process for generating combinatorial libraries with a high percentage of unique and functional mutants,

wherein small groups of residues are randomized in parallel to identify, at each altered position, amino acids which lead to functional proteins. Exponential ensemble mutagenesis is described in Delegrave, S. and Youvan, D.C., Biotechnology Research, 11:1548-1552 (1993), the disclosure of which incorporated herein by reference in its entirety. Random and site-directed mutagenesis are described in Arnold, F.H., Current Opinion in Biotechnology, 4:450-455 (1993), the disclosure of which is incorporated herein by reference in its entirety.

In some embodiments, the variants are created using shuffling procedures wherein portions of a plurality of nucleic acids which encode distinct polypeptides are fused together to create chimeric nucleic acid sequences which encode chimeric polypeptides. Shuffling procedures are described in U.S. Patent Application Serial No. 08/677,112, filed July 9, 1996, U.S. Patent Application Serial No. 08/942,504, filed October 31, 1997, U.S. Patent No. 5,939,250, issued August 17, 1999, and U.S. Patent Application Serial No. 09/375,605, filed August 17, 1999, the disclosures of which are incorporated herein by reference in their entireties.

The variants of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 may be (i) variants in which one or more of the amino acid residues of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code.

Conservative substitutions are those that substitute a given amino acid in a polypeptide by another amino acid of like characteristics. Typically seen as conservative substitutions are the following replacements: replacements of an aliphatic amino acid such as Ala, Val, Leu and Ile with another aliphatic amino acid; replacement of a Ser with a Thr or vice versa; replacement of an acidic residue such as Asp and Glu with another acidic residue; replacement of a residue bearing an amide group, such as Asn and Gln, with another residue bearing an amide group; exchange of a basic residue such as Lys and Arg with another basic residue; and replacement of an aromatic residue such as Phe, Tyr with another aromatic residue.

Other variants are those in which one or more of the amino acid residues of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 includes a substituent group.

Still other variants are those in which the polypeptide is associated with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol).

Additional variants are those in which additional amino acids are fused to the polypeptide, such as a leader sequence, a secretory sequence, a proprotein sequence or a sequence which facilitates purification, enrichment, or stabilization of the polypeptide.

In some embodiments, the fragments, derivatives and analogs retain the same biological function or activity as the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80. In other embodiments, the fragment, derivative, or analog includes a proprotein, such that the fragment, derivative, or analog can be activated by cleavage of the proprotein portion to produce an active polypeptide.

Another aspect of the present invention are polypeptides or fragments thereof which have at least 70%, at least 80%, at least 85%, at least 90%, at least 95%, or more than 95% homology to one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or a fragment comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. Homology may be determined using a program, such as FASTA version 3.0t78 with the default parameters, which aligns the polypeptides or fragments being compared and determines the extent of amino acid identity or similarity between them. It will be appreciated that amino acid "homology" includes conservative amino acid substitutions such as those described above.

The polypeptides or fragments having homology to one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or a fragment comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino

acids thereof may be obtained by isolating the nucleic acids encoding them using the techniques described above.

Alternatively, the homologous polypeptides or fragments may be obtained through biochemical enrichment or purification procedures. The sequence of potentially homologous polypeptides or fragments may be determined by proteolytic digestion, gel electrophoresis and/or microsequencing. The sequence of the prospective homologous polypeptide or fragment can be compared to one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or a fragment comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof using a program such as FASTA version 3.0t78 with the default parameters.

The polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof invention may be used in a variety of applications. For example, the polypeptides or fragments thereof may be used to catalyze biochemical reactions. In particular, the polypeptides of SEQ ID NOs: 14 and 46, which have homology to glutamate semialdehyde amino transferase, or fragments thereof, may be used to catalyze the synthesis of 5-aminolevulinate from S-4-amino-5-oxopentanoate. The polypeptides of SEQ ID NOs: 26 and 58, which have homology to triose phosphate isomerase, or fragments thereof, may be used to catalyze the synthesis of glycerone phosphate from D-glyceraldehyde 3-phosphate. The polypeptides of SEQ ID NOs: 32 and 64, which have homology to dCMP deaminase, or fragments thereof, may be used to catalyze the reaction of deoxyctidine and water to produce deoxyuridine and ammonia. The polypeptides of SEQ ID NOs: 38 and 72, which have homology to the MenA protein, or fragments thereof, may be used to catalyze the synthesis of menaquinone. The polypeptide of SEQ ID NO: 80, which has homology to glucose-1-dehydrogenase, may be used to catalyze the synthesis of D-glucono-1,5-lacctone from D-glucose.

The polypeptide of SEQ ID NO: 10, which has homology to lysyl tRNA synthetase, or fragments thereof, may be used to identify compounds capable of

specifically inhibiting the growth of *Cenarchaeum symbiosis*, since tRNA synthetases are attractive targets for agents which inhibit growth.

Agents which specifically inhibit the activity of the lysyl tRNA synthetase from *Cenarchaeum symbiosum* may be identified using a variety of methods known to those skilled in the art. For example, a plurality of agents may be generated using combinatorial chemistry or recombinant DNA libraries encoding a large number of short peptides. The lysyl tRNA synthetases from *Cenarchaeum symbiosum* and control organisms are contacted with the agents and those agents which bind to the lysyl tRNA synthetase from *Cenarchaeum symbiosum* but not to the enzyme from the control organisms are identified. *Cenarchaeum symbiosum* is then contacted with the identified agents to determine which agents inhibit the organism's growth.

The polypeptides of SEQ ID NOs: 28 and 60, which have homology to the TATA box binding protein, or fragments thereof, may be used to identify promoters in nucleic acids from *Cenarchaeum symbiosis*. In such procedures, the polypeptide or fragment thereof is allowed to contact the nucleic acid and binding of the polypeptide or fragment thereof to the nucleic acid is detected. Binding may be detected by performing a gel shift analysis, a nuclease protection analysis, or by detecting the retention of the nucleic acid on a column matrix having the TATA box binding protein, or a fragment thereof, affixed thereto.

Compounds which specifically inhibit the binding of the TATA box binding protein of *Cenarchaeum symbiosis* to promoters may also be used to inhibit growth of the organism. Such compounds may be identified as described above.

Similarly, agents which specifically inhibit the activity of the polypeptides of SEQ ID NOs: 34 and 66, which have homology to RNA helicase, may be used to inhibit the growth of *Cenarchaeum symbiosis*. Such agents may be identified as described above.

The polypeptides of SEQ ID NOs: 30 and 62, which have homology to DNA polymerase I, or fragments thereof, may be used to insert a detectable label into a nucleic acid or to generate blunt ends on nucleic acids which have been digested with a restriction endonuclease.

The polypeptides of SEQ ID NOs: 42 and 76, which have homology to site specific DNA methyltranseferases, or fragments thereof, may be used in procedures in

which it is desirable to protect nucleic acid sequences from digestion with restriction endonucleases. For example, a nucleic acid sequence having one or more restriction sites therein may be treated with the polypeptides of SEQ ID NOs: 42 or 76 prior to the addition of linkers to the nucleic acid. Thereafter, the linkers may be digested with the restriction enzyme, while the sites in the remainder of the nucleic acid are protected from digestion.

The polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80, or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof, may also be used to generate antibodies which bind specifically to the polypeptides or fragments. The resulting antibodies may be used to determine whether a biological sample contains *Cenarchaeum symbiosum*. In such procedures, a biological sample is contacted with an antibody capable of specifically binding to one of the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. The ability of the biological sample to bind to the antibody is then determined. For example, binding may be determined by labeling the antibody with a detectable label such as a fluorescent agent, an enzymatic label, or a radioisotope. Alternatively, binding of the antibody to the sample may be detected using a secondary antibody having such a detectable label thereon. A variety of assay protocols which may be used to detect the presence of *Cenarchaeum symbiosum* in a sample are familiar to those skilled in the art. Particular assays include ELISA assays, sandwich assays, radioimmunoassays, and Western Blots.

Polyclonal antibodies generated against the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof can be obtained by direct injection of the polypeptides into an animal or by administering the polypeptides to an animal, preferably a nonhuman. The antibody so obtained will then bind the polypeptide itself. In this manner, even a sequence encoding only a fragment of the polypeptide can be used to generate antibodies which may bind to the whole

native polypeptide. Such antibodies can then be used to isolate the polypeptide from cells expressing that polypeptide.

For preparation of monoclonal antibodies, any technique which provides antibodies produced by continuous cell line cultures can be used. Examples include the hybridoma technique (Kohler and Milstein, 1975, Nature, 256:495-497, the disclosure of which is incorporated herein by reference), the trioma technique, the human B-cell hybridoma technique (Kozbor et al., 1983, Immunology Today 4:72, the disclosure of which is incorporated herein by reference), and the EBV-hybridoma technique (Cole, et al., 1985, in Monoclonal Antibodies and Cancer Therapy, Alan R. Liss, Inc., pp. 77-96, the disclosure of which is incorporated herein by reference).

Techniques described for the production of single chain antibodies (U.S. Patent No. 4,946,778, the disclosure of which is incorporated herein by reference) can be adapted to produce single chain antibodies to the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof. Alternatively, transgenic mice may be used to express humanized antibodies to these polypeptides or fragments thereof.

Antibodies generated against the polypeptides of SEQ ID NOs: 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74 76, 78, and 80 or fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids thereof may be used in screening for similar polypeptides from other organisms and samples. In such techniques, polypeptides from the organism are contacted with the antibody and those polypeptides which specifically bind the antibody are detected. Any of the procedures described above may be used to detect antibody binding. One such screening assay is described in "Methods for Measuring Cellulase Activities", *Methods in Enzymology*, Vol 160, pp. 87-116, which is hereby incorporated by reference in its entirety.

As used herein the term "nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77" encompasses the nucleotide sequences of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75,

79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77, fragments of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77, nucleotide sequences homologous to SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or homologous to fragments of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77, and sequences complementary to all of the preceding sequences. The fragments include portions of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 comprising at least 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, or 500 consecutive nucleotides of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77. Preferably, the fragments are novel fragments. Homologous sequences and fragments of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 refer to a sequence having at least 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75% or 70% homology to these sequences. Homology may be determined using any of the computer programs and parameters described herein, including BLASTN version 2.0 with the default parameters. Homologous sequences also include RNA sequences in which uridines replace the thymines in the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77. The homologous sequences may be obtained using any of the procedures described herein or may result from the correction of a sequencing error. It will be appreciated that the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 can be represented in the traditional single character format (See the inside back cover of Stryer, Lubert. *Biochemistry*, 3rd edition. W. H Freeman & Co., New York.) or in any other format which records the identity of the nucleotides in a sequence.

As used herein the term "polypeptide codes of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36,

40, 44, 48, 50, 52, 54, 56, 70, 74, and 78" encompasses the polypeptide sequence of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 which are encoded by the extended cDNAs of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77, polypeptide sequences homologous to the polypeptides of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78, or fragments of any of the preceding sequences. Homologous polypeptide sequences refer to a polypeptide sequence having at least 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75% or 70% homology to one of the polypeptide sequences of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. Homology may be determined using any of the computer programs and parameters described herein, including FASTA version 3.0t78 with the default parameters or with any modified parameters. The homologous sequences may be obtained using any of the procedures described herein or may result from the correction of a sequencing error. The polypeptide fragments comprise at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of the polypeptides of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. Preferably, the fragments are novel fragments. It will be appreciated that the polypeptide codes of the SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 can be represented in the traditional single character format or three letter format (See the inside back cover of Starrier, Lubert. *Biochemistry*, 3rd edition. W. H Freeman & Co., New York.) or in any other format which relates the identity of the polypeptides in a sequence.

It will be appreciated by those skilled in the art that the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 and polypeptide codes of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 can be stored, recorded, and manipulated on any medium which can be read and accessed by a

computer. As used herein, the words "recorded" and "stored" refer to a process for storing information on a computer medium. A skilled artisan can readily adopt any of the presently known methods for recording information on a computer readable medium to generate manufactures comprising one or more of the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77, one or more of the polypeptide codes of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, 15, or 20 nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77.

Another aspect of the invention is a computer readable medium having recorded thereon one or more of the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, and 79. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, or 15 of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, and 79.

Another aspect of the present invention is a computer readable medium having recorded thereon one or more of the nucleic acid codes of SEQ ID NOs. 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, or 15 of SEQ ID NOs. 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77.

Another aspect of the present invention is a computer readable medium having recorded thereon one or more of the polypeptide codes of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. Another aspect of the present invention is a computer readable medium having recorded thereon one or more of the the polypeptide codes of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, and 80. Another aspect of the present invention is a computer readable medium having recorded thereon one or more of the the polypeptide codes of SEQ ID NOS. 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, 15, or 20 polypeptide codes of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, or 15 polypeptide codes of SEQ ID NOS. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, and 80. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, or 15 polypeptide codes of SEQ ID NOS. 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

Computer readable media include magnetically readable media, optically readable media, electronically readable media and magnetic/optical media. For example, the computer readable media may be a hard disk, a floppy disk, a magnetic tape, CD-ROM, Digital Versatile Disk (DVD), Random Access Memory (RAM), or Read Only Memory (ROM) as well as other types of other media known to those skilled in the art.

Embodiments of the present invention include systems, particularly computer systems which store and manipulate the sequence information described herein. One example of a computer system 100 is illustrated in block diagram form in Figure 3. As used herein, "a computer system" refers to the hardware components, software components, and data storage components used to analyze the nucleotide sequences of the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the sequences of the polypeptide codes of 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. The computer system 100 preferably includes a processor for processing, accessing and manipulating the sequence data. The processor 105 can be any well-known type of central processing unit, such as the Pentium III from Intel Corporation, or similar processor from Sun, Motorola, Compaq or International Business Machines.

Preferably, the computer system 100 is a general purpose system that comprises the processor 105 and one or more internal data storage components 110 for storing data, and one or more data retrieving devices for retrieving the data stored on the data storage components. A skilled artisan can readily appreciate that any one of the currently available computer systems are suitable.

In one particular embodiment, the computer system 100 includes a processor 105 connected to a bus which is connected to a main memory 115 (preferably implemented as RAM) and one or more internal data storage devices 110, such as a hard drive and/or other computer readable media having data recorded thereon. In some embodiments, the computer system 100 further includes one or more data retrieving device 118 for reading the data stored on the internal data storage devices 110.

The data retrieving device 118 may represent, for example, a floppy disk drive, a compact disk drive, a magnetic tape drive, etc. In some embodiments, the internal data storage device 110 is a removable computer readable medium such as a floppy disk, a compact disk, a magnetic tape, etc. containing control logic and/or data recorded thereon. The computer system 100 may advantageously include or be programmed by appropriate software for reading the control logic and/or the data from the data storage component once inserted in the data retrieving device.

The computer system 100 includes a display 120 which is used to display output to a computer user. It should also be noted that the computer system 100 can be linked to other computer systems 125a-c in a network or wide area network to provide centralized access to the computer system 100.

Software for accessing and processing the nucleotide sequences of the nucleic acid codes of SEQ ID Nos.1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 (such as search tools, compare tools, and modeling tools etc.) may reside in main memory 115 during execution.

In some embodiments, the computer system 100 may further comprise a sequence comparer for comparing the above-described nucleic acid codes of SEQ ID Nos. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 stored on a computer readable medium to reference nucleotide or polypeptide sequences stored on a computer readable medium. A "sequence comparer" refers to one or more programs which are

implemented on the computer system 100 to compare a nucleotide sequence with other nucleotide sequences and/or compounds stored within the data storage means. For example, the sequence comparer may compare the nucleotide sequences of the nucleic acid codes of SEQ ID Nos. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 stored on a computer readable medium to reference sequences stored on a computer readable medium to identify homologies or structural motifs. Various sequence comparer programs identified elsewhere in this patent specification are particularly contemplated for use in this aspect of the invention. Protein and/or nucleic acid sequence homologies may be evaluated using any of the variety of sequence comparison algorithms and programs known in the art. Such algorithms and programs include, but are by no means limited to, TBLASTN, BLASTN, BLASTP, FASTA, TFASTA, and CLUSTALW (Pearson and Lipman, 1988, *Proc. Natl. Acad. Sci. USA 85(8)*:2444-2448; Altschul *et al.*, 1990, *J. Mol. Biol. 215(3)*:403-410; Thompson *et al.*, 1994, *Nucleic Acids Res. 22(2)*:4673-4680; Higgins *et al.*, 1996, *Methods Enzymol. 266*:383-402; Altschul *et al.*, 1990, *J. Mol. Biol. 215(3)*:403-410; Altschul *et al.*, 1993, *Nature Genetics 3*:266-272).

In one embodiment, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST") which is well known in the art (see, *e.g.*, Karlin and Altschul, 1990, *Proc. Natl. Acad. Sci. USA 87*:2267-2268; Altschul *et al.*, 1990, *J. Mol. Biol. 215*:403-410; Altschul *et al.*, 1993, *Nature Genetics 3*:266-272; Altschul *et al.*, 1997, *Nuc. Acids Res. 25*:3389-3402). In particular, five specific BLAST programs are used to perform the following task:

(1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database;

(2) BLASTN compares a nucleotide query sequence against a nucleotide sequence database;

(3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database;

(4) TBLASTN compares a query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and

(5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (*i.e.*, aligned) by means of a scoring matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet *et al.*, 1992, *Science 256*:1443-1445; Henikoff and Henikoff, 1993, *Proteins 17*:49-61). Less preferably, the PAM or PAM250 matrices may also be used (see, *e.g.*, Schwartz and Dayhoff, eds., 1978, *Matrices for Detecting Distance Relationships: Atlas of Protein Sequence and Structure*, Washington: National Biomedical Research Foundation). BLAST programs are accessible through the U.S. National Library of Medicine, e.g., at www.ncbi.nlm.nih.gov.

The BLAST programs evaluate the statistical significance of all high-scoring segment pairs identified, and preferably selects those segments which satisfy a user-specified threshold of significance, such as a user-specified percent homology. Preferably, the statistical significance of a high-scoring segment pair is evaluated using the statistical significance formula of Karlin (see, *e.g.*, Karlin and Altschul, 1990, *Proc. Natl. Acad. Sci. USA 87*:2267-2268).

The parameters used with the above algorithms may be adapted depending on the sequence length and degree of homology studied. In some embodiments, the parameters may be the default parameters used by the algorithms in the absence of instructions from the user.

Figure 4 is a flow diagram illustrating one embodiment of a process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the homology levels between the new sequence and the sequences in the

database. The database of sequences can be a private database stored within the computer system 100, or a public database such as GENBANK that is available through the Internet.

The process 200 begins at a start state 201 and then moves to a state 202 wherein the new sequence to be compared is stored to a memory in a computer system 100. As discussed above, the memory could be any type of memory, including RAM or an internal storage device.

The process 200 then moves to a state 204 wherein a database of sequences is opened for analysis and comparison. The process 200 then moves to a state 206 wherein the first sequence stored in the database is read into a memory on the computer. A comparison is then performed at a state 210 to determine if the first sequence is the same as the second sequence. It is important to note that this step is not limited to performing an exact comparison between the new sequence and the first sequence in the database. Well-known methods are known to those of skill in the art for comparing two nucleotide or protein sequences, even if they are not identical. For example, gaps can be introduced into one sequence in order to raise the homology level between the two tested sequences. The parameters that control whether gaps or other features are introduced into a sequence during comparison are normally entered by the user of the computer system.

Once a comparison of the two sequences has been performed at the state 210, a determination is made at a decision state 210 whether the two sequences are the same. Of course, the term "same" is not limited to sequences that are absolutely identical. Sequences that are within the homology parameters entered by the user will be marked as "same" in the process 200.

If a determination is made that the two sequences are the same, the process 200 moves to a state 214 wherein the name of the sequence from the database is displayed to the user. This state notifies the user that the sequence with the displayed name fulfills the homology constraints that were entered. Once the name of the stored sequence is displayed to the user, the process 200 moves to a decision state 218 wherein a determination is made whether more sequences exist in the database. If no more sequences exist in the database, then the process 200 terminates at an end state 220. However, if more sequences do exist in the database, then the process 200 moves to a state 224 wherein a pointer is moved to the next sequence in the database so that it can be

compared to the new sequence. In this manner, the new sequence is aligned and compared with every sequence in the database.

It should be noted that if a determination had been made at the decision state 212 that the sequences were not homologous, then the process 200 would move immediately to the decision state 218 in order to determine if any other sequences were available in the database for comparison.

Accordingly, one aspect of the present invention is a computer system comprising a processor, a data storage device having stored thereon a nucleic acid code of SEQ ID Nos. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78, a data storage device having retrievably stored thereon reference nucleotide sequences or polypeptide sequences to be compared to the nucleic acid code of SEQ ID Nos.1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 and a sequence comparer for conducting the comparison. The sequence comparer may indicate a homology level between the sequences compared or identify structural motifs in the above described nucleic acid code of SEQ ID Nos. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 or it may identify structural motifs in sequences which are compared to these nucleic acid codes and polypeptide codes. In some embodiments, the data storage device may have stored thereon the sequences of at least 2, 5, 10, 15, 20, 25, 30 or 40 or more of the nucleic acid codes of SEQ ID Nos. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

Another aspect of the present invention is a method for determining the level of homology between a nucleic acid code of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 and a reference nucleotide sequence or polypeptide sequence, comprising the steps of reading the nucleic acid code or the polypeptide code and the reference nucleotide or polypeptide sequence through the use of a computer program which determines homology levels and determining homology between the nucleic acid code or polypeptide code and the reference nucleotide or polypeptide sequence with the computer program. The computer program may be any of a number of computer programs for determining homology levels, including those specifically enumerated herein, including BLAST2N or BLASTN with the default parameters or with any modified parameters. The method may be implemented using the computer systems described above. The method may also be performed by reading at least 2, 5, 10, 15, 20, 25, 30 or 40 or more of the above described nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 through use of the computer program and determining homology between the nucleic acid codes or polypeptide codes and reference nucleotide sequences or polypeptide sequences.

Figure 5 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous. The process 250 begins at a start state 252 and then moves to a state 254 wherein a first sequence to be compared is stored to a memory. The second sequence to be compared is then stored to a memory at a state 256. The process 250 then moves to a state 260 wherein the first character in the first sequence is read and then to a state 262 wherein the first character of the second sequence is read. It should be understood that if the sequence is a nucleotide sequence, then the character would normally be either A, T, C, G or U. If the sequence is a protein sequence, then it is preferably in the single letter amino acid code so that the first and sequence sequences can be easily compared.

A determination is then made at a decision state 264 whether the two characters are the same. If they are the same, then the process 250 moves to a state 268 wherein the next characters in the first and second sequences are read. A determination is then made whether the next characters are the same. If they are, then the process 250 continues this loop until two characters are not the same. If a determination is made that the next two characters are not the same, the process 250 moves to a decision state 274 to determine whether there are any more characters either sequence to read.

If there aren't any more characters to read, then the process 250 moves to a state 276 wherein the level of homology between the first and second sequences is displayed to the user. The level of homology is determined by calculating the proportion of characters between the sequences that were the same out of the total number of sequences in the first sequence. Thus, if every character in a first 100 nucleotide sequence aligned with a every character in a second sequence, the homology level would be 100%.

Alternatively, the computer program may be a computer program which compares the nucleotide sequences of the nucleic acid codes of the present invention, to reference nucleotide sequences in order to determine whether the nucleic acid code of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 differs from a reference nucleic acid sequence at one or more positions. Optionally such a program records the length and identity of inserted, deleted or substituted nucleotides with respect to the sequence of either the reference polynucleotide or the nucleic acid code of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77. In one embodiment, the computer program may be a program which determines whether the nucleotide sequences of the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 contain a single nucleotide polymorphism (SNP) with respect to a reference nucleotide sequence.

Accordingly, another aspect of the present invention is a method for determining whether a nucleic acid code of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51,

53, 55, 69, 73 and 77 differs at one or more nucleotides from a reference nucleotide sequence comprising the steps of reading the nucleic acid code and the reference nucleotide sequence through use of a computer program which identifies differences between nucleic acid sequences and identifying differences between the nucleic acid code and the reference nucleotide sequence with the computer program. In some embodiments, the computer program is a program which identifies single nucleotide polymorphisms. The method may be implemented by the computer systems described above and the method illustrated in Figure 6. The method may also be performed by reading at least 2, 5, 10, 15, 20, 25, 30, or 40 of the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 and the reference nucleotide sequences through the use of the computer program and identifying differences between the nucleic acid codes and the reference nucleotide sequences with the computer program.

In other embodiments the computer based system may further comprise an identifier for identifying features within the nucleotide sequences of the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

An "identifier" refers to one or more programs which identifies certain features within the above-described nucleotide sequences of the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. In one embodiment, the identifier may comprise a program which identifies an open reading frame in the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77.

Figure 7 is a flow diagram illustrating one embodiment of an identifier process 300 for detecting the presence of a feature in a sequence. The process 300 begins at a start state 302 and then moves to a state 304 wherein a first sequence that is to be checked for features is stored to a memory 115 in the computer system 100. The process 300 then moves to a state 306 wherein a database of sequence features is opened. Such a database would include a list of each feature's attributes along with the name of the feature. For example, a feature name could be "Initiation Codon" and the attribute would be "ATG". Another example would be the feature name "TAATAA Box" and the feature attribute would be "TAATAA". An example of such a database is produced by the University of Wisconsin Genetics Computer Group (www.gcg.com). Alternatively, the features may be structural polypeptide motifs such as alpha helices, beta sheets, or functional polypeptide motifs such as enzymatic active sites, helix-turn-helix motifs or other motifs known to those skilled in the art.

Once the database of features is opened at the state 306, the process 300 moves to a state 308 wherein the first feature is read from the database. A comparison of the attribute of the first feature with the first sequence is then made at a state 310. A determination is then made at a decision state 316 whether the attribute of the feature was found in the first sequence. If the attribute was found, then the process 300 moves to a state 318 wherein the name of the found feature is displayed to the user.

The process 300 then moves to a decision state 320 wherein a determination is made whether move features exist in the database. If no more features do exist, then the process 300 terminates at an end state 324. However, if more features do exist in the database, then the process 300 reads the next sequence feature at a state 326 and loops back to the state 310 wherein the attribute of the next feature is compared against the first sequence.

It should be noted, that if the feature attribute is not found in the first sequence at the decision state 316, the process 300 moves directly to the decision state 320 in order to determine if any more features exist in the database.

Accordingly, another aspect of the present invention is a method of identifying a feature within the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28,

30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 comprising reading the nucleic acid code(s) or polypeptide code(s) through the use of a computer program which identifies features therein and identifying features within the nucleic acid code(s) with the computer program. In one embodiment, computer program comprises a computer program which identifies open reading frames. The method may be performed by reading a single sequence or at least 2, 5, 10, 15, 20, 25, 30, or 40 of the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 through the use of the computer program and identifying features within the nucleic acid codes or polypeptide codes with the computer program.

The nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 may be stored and manipulated in a variety of data processor programs in a variety of formats. For example, the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78 may be stored as text in a word processing file, such as MicrosoftWORD or WORDPERFECT or as an ASCII file in a variety of database programs familiar to those of skill in the art, such as DB2, SYBASE, or ORACLE. In addition, many computer programs and databases may be used as sequence comparers, identifiers, or sources of reference nucleotide sequences or polypeptide sequences to be compared to the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78. The

following list is intended not to limit the invention but to provide guidance to programs and databases which are useful with the nucleic acid codes of SEQ ID NOs. 1, 2, 5, 9, 13, 25, 27, 29, 31, 33, 37, 41, 45, 57, 59, 61, 63, 65, 67, 71, 75, 79, 3, 7, 11, 15, 17, 19, 21, 23, 35, 39, 43, 47, 49, 51, 53, 55, 69, 73 and 77 or the polypeptide codes of SEQ ID NOs. 6, 10, 14, 26, 28, 30, 32, 34, 38, 42, 46, 58, 60, 62, 64, 66, 68, 72, 76, 80, 4, 8, 12, 16, 18, 20, 22, 24, 36, 40, 44, 48, 50, 52, 54, 56, 70, 74, and 78.

The programs and databases which may be used include, but are not limited to: MacPattern (EMBL), DiscoveryBase (Molecular Applications Group), GeneMine (Molecular Applications Group), Look (Molecular Applications Group), MacLook (Molecular Applications Group), BLAST and BLAST2 (NCBI), BLASTN and BLASTX (Altschul et al, *J. Mol. Biol.* 215: 403 (1990)), FASTA (Pearson and Lipman, *Proc. Natl. Acad. Sci. USA,* 85: 2444 (1988)), FASTDB (Brutlag et al. Comp. App. Biosci. 6:237-245, 1990), Catalyst (Molecular Simulations Inc.), Catalyst/SHAPE (Molecular Simulations Inc.), Cerius^2.DBAccess (Molecular Simulations Inc.), HypoGen (Molecular Simulations Inc.), Insight II, (Molecular Simulations Inc.), Discover (Molecular Simulations Inc.), CHARMm (Molecular Simulations Inc.), Felix (Molecular Simulations Inc.), DelPhi, (Molecular Simulations Inc.), QuanteMM, (Molecular Simulations Inc.), Homology (Molecular Simulations Inc.), Modeler (Molecular Simulations Inc.), ISIS (Molecular Simulations Inc.), Quanta/Protein Design (Molecular Simulations Inc.), WebLab (Molecular Simulations Inc.), WebLab Diversity Explorer (Molecular Simulations Inc.), Gene Explorer (Molecular Simulations Inc.), SeqFold (Molecular Simulations Inc.), the MDL Available Chemicals Directory database, the MDL Drug Data Report data base, the Comprehensive Medicinal Chemistry database, Derwents's World Drug Index database, the BioByteMasterFile database, the Genbank database, and the Genseqn database. Many other programs and data bases would be apparent to one of skill in the art given the present disclosure.

Motifs which may be detected using the above programs include sequences encoding leucine zippers, helix-turn-helix motifs, glycosylation sites, ubiquitination sites, alpha helices, and beta sheets, signal sequences encoding signal peptides which direct the secretion of the encoded proteins, sequences implicated in transcription regulation such as homeoboxes, acidic stretches, enzymatic active sites, substrate binding sites, and enzymatic cleavage sites.

The present invention will be further described with reference to the following examples; however, it is to be understood that the present invention is not limited to such examples.

In order to begin the physiological characterization of *Cenarchaeum symbiosum*, it was necessary to obtain enriched preparations of *Cenarchaeum symbiosum* for use in the construction of genomic DNA libraries in fosmid based vectors. Genomic DNA libraries were constructed from two enriched preparations using the methods described in Example 1 below.

Example 1

Enrichment of *Cenarchaeum symbiosum* Cells in Samples Obtained from *Axinella Mexicana*

Enriched preparations of *Cenarchaeum symbiosum* for use in the preparation of the first fosmid genomic DNA library were obtained essentially as described in Preston, C. M. *et al.* 1996. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc. Natl. Acad. Sci.* USA **93,** 6241-6246, the disclosure of which is incorporated herein by reference. Briefly, a small individual of *A. mexicana* was incubated in calcium- and magnesium-free artificial seawater (ASW) containing 0.25 mg/ml Pronase. The tissue was then homogenized and enriched for archaeal cells by differential centrifugation.

Enriched preparations of *Cenarchaeum symbiosum* for use in preparing the second fosmid genomic DNA library were obtained from a different sponge individual using the following improved enrichment procedure. A small individual of *A. mexicana* was incubated in calcium- and magnesium-free artificial seawater (460mm NaCl, 11mM KCl, 7mM Na_2SO_4, 2mM $NaHCO_3$) containing 0.25 mg/ml Pronase at room temperature for one hour. The sponge tissue was rinsed in artificial seawater and homogenized in a blender. Large particles and spicules were removed by low-speed centrifugation (4000 rpm, Sorvall GSA rotor at 4°C). The supernatant was next centrifuged at 5000 rpm for 5 min. at 4°C to remove large sponge cells, and the resulting supernatant was centrifuged at 10,000 rpm in a GSA rotor at 4°C for 20 min. to collect the *Cenarchaeum symbiosum* cells. Following centrifugation, the recovered cell fraction containing *Cenarchaeum symbiosum* was further incubated for 1 hr at 4°C

in 10 mM Tris/HCl pH 8 and 200 mM EDTA. The cells were then pelleted and subsequently purified on a 15 % Percoll (Sigma) cushion in artificial sea water centrifuged at 2500 rpm in a Beckman SS34 rotor. Archaeal cells banded in the light, upper fraction after centrifugation. This cell fraction was washed in ASW and resuspended in TE buffer (10 mM TrisHCl pH 8, 0.1 mM EDTA). The additional incubation step was found to increase the lysis of sponge cells, which resulted in an enhanced separation of archaeal and eukaryotic cells in the percoll gradient.

Quantitative hybridization experiments were performed as described in DeLong, E. F. 1992. Archaea in coastal marine environments. *Proc. Natl. Acad. Sci.* **89**, 5685-5689, the disclosure of which is incorporated herein by reference, using an oligonucleotide specific for archaea having the sequence GTGCTCCCCCGCCAATTCCT (SEQ ID NO: 115). These hybridization experiments indicated that 25% to 30% of the total rRNA from this fraction was derived from archaea.

The enriched cell preparations were then utilized to construct fosmid libraries as described in Example 2 below.

Example 2

Construction of Fosmid Libraries

DNA was extracted from the enriched preparations of Example 1 and inserted into fosmids as described in Preston, C. M. *et al.* 1996. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc. Natl. Acad. Sci.* USA **93,** 6241-6246 and Stein, J.L. *et al.* 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.* **178**, 591-599, the disclosures of which are incorporated herein by reference. A vertical cross section of sponge (0.5 g) was mechanically dissociated in 0.22μm filtered, autoclaved seawater using a tissue homogenizer. Cell lysis was accomplished by incubating the dissociated cells in 1 mg of lysozyme per ml for 30 min. at 37°C followed by an incubation for 30 min. at 55°C with 0.5mg of proteinase K per ml and 1% SDS. The tubes were finally placed in a boiling water bath for 60 sec to complete lysis. The protein fraction was removed with two extractions with phenol:chloroform:isoamyl alcohol (50:49:1), pH 8.0, followed by

a chloroform: isoamyl alcohol (24:1) extraction. Nucleic acids were ethanol-precipitated and resuspended in TE buffer (10mM Tris.HCl/1mM Na_2-EDTA, pH 8.0). Approximately 5μg of DNA was purified by CsCl equilibrium density gradiant ultracentriguation on a Beckman Optima tabletop ultracentrifuge using a TLA100 rotor.

The genomic DNA obtained above was inserted into fosmids as follows. The genomic DNA was partially digested with Sau3AI (Promega) and treated with heat-labile phosphatase (HK phosphatase; Epicentre). The partially digested genomic DNA was ligated with pFOS (*See* U.J. Kim et al., Nucleic Acids Res. **20:**1083-1085 (1992), the disclosure of which is incorporated herein by reference) which had previously been digested with AatII, phosphatase treated (HK phosphatase), and subsequently digested with BamHI. The ligation mixture was used for *in vitro* packaging with the Gigapack XL packaging system (Stratagene) selecting for DNA inserts of 35 to 45kb. The phage particles were transfected into *E. coli* DH10B (Bethesda Research LaboratoriesP and the cells were spread onto LB plates supplemented with 12.5μg/ml chloramphenicol.

Example 3

Identification of Fosmids Containing the *Cenarchaeum symbiosum* rRNA Operon

The fosmid libraries constructed above were screened to identify clones containing the rRNA operon. PCR reactions were conducted on the library using primers known to amplify the rRNA operon.

The first fosmid library yielded seven unique clones, out of a total of 10,236 recombinant fosmids, which contained the *Cenarchaeum symbiosum* rRNA operon. The second fosmid library yielded eight unique clones, out of a total of 2100 recombinant fosmids, which contained the *Cenarchaeum symbiosum* rRNA operon.

The sequences of the 16S rRNA genes in each of the 15 fosmids containing the *Cenarchaeum symbiosum* rRNA operon were determined. The sequences of the small subunit rRNA genes of these 15 fosmids exhibited variations with respect to one another. Ten of the fosmids contained a small subunit rRNA gene having the sequence of the 16S rRNA gene in the insert of SEQ ID NO: 1, while the remaining fosmids contained a small subunit rRNA gene having the sequence of the 16S rRNA gene in the insert of SEQ ID NO: 2. As discussed in more detail below, the differences in the

sequences of the rRNA genes may be used to determine whether a sample contains *Cenarchaeum symbiosum* variant A or *Cenarchaeum symbiosum* variant B.

In addition to determining the sequences of the rRNA genes, the sequences adjacent to the rRNA genes were also determined.

Example 4

Fosmid Sequencing

Partial restriction enzyme digests were conducted on two purified fosmids, fosmid 101G10 (which contains the variant A sequence) and fosmid 60A5 (which contains the variant B sequence). The partially digested DNA was used to construct plasmid libraries containing inserts of 1-2 kb. The resulting plasmids were sequenced using Applied Biosystems (ABI, Foster City, CA) Prism Dye-terminator FS reaction mix. Direct sequencing from fosmids was used for gap filling and resequencing to ensure accuracy. Fosmid sequencing was performed by using DNA from a single 3 ml overnight culture purified on an Autogen 740 automated plasmid isolation system. Each reaction consisted of one preparation of DNA directly resuspended by the addition of 16 µl H_2O, 8µl oligonucleotide primer (1.4 pmol/µl) and 16 µl ABI Prism Dye-terminator FS reaction mix. Cycle sequencing was performed with a 96° C 3 min. preincubation followed by 25 cycles of the sequence 96° C 20 sec. / 50° C 20 sec. / 60° C 4 min. and a 5 min. post-cycling incubation at 60° C. Sequencing reaction products were analyzed on ABI 377 Prism Sequencers.

The complete sequences of the *Cenarchaeum symbiosum* derived inserts in the two fosmids are provided in the accompanying sequence listing as SEQ ID NO: 1 (fosmid 101G10) and SEQ ID NO: 2 (fosmid 60A5). The insert of fosmid 101G10 (SEQ ID NO: 1, designated variant A) was 32,998 bp and was syntenic over ca. 28 kbp with the 42,432 bp insert of fosmid 60A5 (SEQ ID NO:2, designated variant B). Analysis of the common 28 kbp region is shown in Fig. 1.

Although the sequences of both fosmids could be aligned unambiguously over most of the overlapping region, four large insertion/deletions ranging in size from 142 bp to 1994 bp were identified between positions 20,500 and 25,800. The longest insertion contained a repetitive element of 1784 bp, that was found in the sequence of SEQ ID NO: 1 between *men*A and ORF05. It was composed of a 3-fold direct repeat of

575 bp (rep1 through 3 in Fig. 1), with repeats exhibiting only minor sequence variation (95.8% to 98.7% identity).

A segment of 56 bp at the start of this repeat was also found adjacent to the 3' terminus of the third direct repeat. No obvious structural or sequence similarities to known repeats or mobile genetic elements from other organisms were identified within the repeat sequence. Its occurrence in only one variant and its relatively low G+C content relative to the rest of the fragment suggest that it may have been acquired by horizontal transfer from a different genetic context.

The sequenced regions contained several open reading frames or RNA encoding sequences. Some of the identified open reading frames encode proteins having homology to previously identified proteins. In particular, some of the open reading frames encode proteins involved in several metabolic pathways, providing insight into the physiology of *Cenarchaeum symbiosum.*

An open reading frame which encodes a protein having homology to glutamate semialdehyde aminotransferase (a protein involved in heme biosynthesis) was identified between nucleotides 7604-8908 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 23558-24682 of the insert from fosmid 60A5 (SEQ ID NO: 2) . These open reading frames have been assigned SEQ ID NOs: 45 and 13 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 46 and 14 respectively in the accompanying sequence listing. A gene encoding glutamate semialdehyde aminotransferase has also been detected in a rRNA operon containing genomic fragment of a planktonic marine crenarchaeote. (Stein, J.L. *et al.* 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.* **178**, 591-599)

An open reading frame encoding a protein having homology to triose-phosphate isomerase was identified between 13944-14612 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 29655-30491 of the insert from fosmid 60A5 (SEQ ID NO: 2) . These open reading frames have been assigned SEQ ID NOs: 57 and 25 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 58 and 26 respectively in the accompanying sequence

listing. This triosephosphate isomerase represents the first such protein sequence reported in a crenarchaeote, and shares known archaeal signature sequences and deletions which distinguish archaeal triosephosphate isomerase genes from their eucaryal and eubacterial homologues.

An open reading frame encoding a protein having homology to the TATA binding protein was identified between 14616-15164 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 30501-31049 of the insert from fosmid 60A5 (SEQ ID NO: 2) on the strands complementary to the insert strands provided in SEQ ID NOs: 1 and 2. These open reading frames have been assigned SEQ ID NOs: 59 and 27 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 60 and 28 respectively in the companying sequence listing. This TATA box-binding protein (TBP) is similar to other known archaeal TBP's and is N-terminally truncated with respect to the eukaryal homologs. It shares 49% amino acid similarity with TBP from *Pyrococcus woesii*.

An open reading frame encoding a protein having homology to DNA polymerase (a protein involved in DNA replication and repair) was identified between nucleotides 15488-18025 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 31371-33905 of the insert from fosmid 60A5 (SEQ ID NO: 2) on the strands complementary to the insert strands provided in SEQ ID NOs: 1 and 2. These open reading frames have been assigned SEQ ID NOs: 61 and 29 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 62 and 30 respectively in the accompanying sequence listing.

The DNA polymerase of *Cenarchaeum symbiosum* has a high degree of similarity to the crenarchaeal homologs from the extreme thermophiles *Sulfolobus acidocaldarius* and *Pyrodictium occultum* (54% and 53% resp.) and exhibits all conserved motifs of B-(a-)type DNA polymerases and 3'-5'-exonuclease motifs, both indicative of archaeal polymerases. A more detailed phylogenetic analysis and biochemical characterization of the *C. symbiosum* polymerase has been published elsewhere. (Schleper, C., *et al.* 1997. Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon *Cenarchaeum symbiosum*. *J. Bact.* 179, 7803-7811)

An open reading frame which encodes a protein having homology to dCMP deaminase (a protein involved in pyrimidine synthesis) was identified between nucleotides 18022-18663 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 33902-34456 of the insert from fosmid 60A5 (SEQ ID NO: 2) on the strands complementary to the insert strands provided in SEQ ID NOs: 1 and 2. These open reading frames have been assigned SEQ ID NOs: 63 and 31 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 64 and 32 respectively in the accompanying sequence listing.

An open reading frame encoding a protein having homology to the ATP dependent RNA helicase (a protein involved in translation) was identified between nucleotides 18638-20149 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 34559-36067 of the insert from fosmid 60A5 (SEQ ID NO: 2) . These open reading frames have been assigned SEQ ID NOs: 65 and 33 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 66 and 34 respectively in the accompanying sequence listing. The identified ATP RNA helicase is highly similar in sequence to homologues found in the genomic sequences of three euryarchaeota (Bult, C., *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058-1073; Klenk, H.P. *et al.* 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364-370; Smith, D. R.*et al.* 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135-7155).

An open reading frame encoding a protein having homology to MenA (a protein involved in menaquinone biosynthesis) was identified between nucleotides 20956-21834 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 37404-38282 of the insert from fosmid 60A5 (SEQ ID NO: 2). These open reading frames have been assigned SEQ ID NOs: 71 and 37 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 72 and 38 respectively in the accompanying sequence listing.

An open reading frame encoding a protein having homology to the site specific DNA methyltranseferase proteins involved in restriction/modification was identified between nucleotides 26378-27454 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 40563-41669 of the insert from fosmid 60A5 (SEQ ID NO: 2) on the strands complementary to the insert strands provided in SEQ ID NOs: 1 and 2. These open reading frames have been assigned SEQ ID NOs: 75 and 41 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 76 and 42 respectively in the accompanying sequence listing.

An open reading frame encoding a protein having homology to the histone H1 DNA binding protein was identified between nucleotides 10625-1134 of the insert from fosmid 60A5 (SEQ ID NO: 2) . This open reading frame has been assigned SEQ ID No: 5 in the accompanying sequence listing, while the polypeptide it encodes has been assigned SEQ ID No: 6 in the accompanying sequence listing.

An open reading frame encoding a protein having homology to lysyl tRNA synthetase was identified between nucleotides 13046-14620 of the insert from fosmid 60A5 (SEQ ID NO: 2). This open reading frame has been assigned SEQ ID No: 9 in the accompanying sequence listing, while the polypeptide it encodes has been assigned SEQ ID No: 10 in the accompanying sequence listing.

A hypothetical open reading frame was identified between nucleotides 11478-13046 of the insert from fosmid 60A5 (SEQ ID NO: 2). This open reading frame has been assigned SEQ ID No: 7 in the accompanying sequence listing, while the polypeptide it encodes has been assigned SEQ ID No: 8 in the accompanying sequence listing.

An open reading frame encoding a protein having homology to peptidylprolyl cis/trans isomerase (a chaperone) was identified between nucleotides 20156-20434 of the insert from fosmid 101G10 (SEQ ID NO: 1) on the strand complementary to that provided in the sequence listing. This open reading frame has been assigned SEQ ID No: 67 in the accompanying sequence listing, while the polypeptide it encodes has been assigned SEQ ID No: 68 in the accompanying sequence listing.

An open reading frame encoding a protein having homology to glucose-1-dehydrogenase was identified between nucleotides 28065-29843 of the insert from

fosmid 101G10 (SEQ ID NO: 1) . This open reading frame has been assigned SEQ ID No: 79 in the accompanying sequence listing, while the polypeptide it encodes has been assigned SEQ ID No: 80 in the accompanying sequence listing.

A hypothetical open reading frame designated Hypothetical 01 was identified between nucleotides 1358-2290 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 17329-18213 of the insert from fosmid 60A5 (SEQ ID NO: 2) on the strands complementary to the insert strands provided in SEQ ID NOs: 1 and 2. These open reading frames have been assigned SEQ ID NOs: 43 and 11 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 44 and 12 respectively in the accompanying sequence listing.

A hypothetical open reading frame designated Hypothetical 02 was identified between nucleotides 8961-9767 of the insert from fosmid 101G10 (SEQ ID NO: 1) between nucleotides 24913-25728 of the insert from fosmid 60A5 (SEQ ID NO: 2). These open reading frames have been assigned SEQ ID NOs: 47 and 15 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 48 and 16 respectively in the accompanying sequence listing.

An open reading frame designated ORF 01 was identified between nucleotides 9772-10479 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 25732-26427 of the insert from fosmid 60A5 (SEQ ID NO: 2) on the strands complementary to the insert strands provided in SEQ ID NOs: 1 and 2. These open reading frames have been assigned SEQ ID NOs: 49 and 17 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 50 and 18 respectively in the accompanying sequence listing.

An open reading frame designated ORF 02 was identified between nucleotides 10545-10922 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 26504-26881 of the insert from fosmid 60A5 (SEQ ID NO: 2). These open reading frames have been assigned SEQ ID NOs: 51 and 19 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 52 and 20 respectively in the accompanying sequence listing.

An open reading frame designated ORF 03 was identified between nucleotides 11382-11987 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between

nucleotides 27337-27936 of the insert from fosmid 60A5 (SEQ ID NO: 2) on the strands complementary to the insert strands provided in SEQ ID NOs: 1 and 2. These open reading frames have been assigned SEQ ID NOs: 53 and 21 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 54 and 22 respectively in the accompanying sequence listing.

An open reading frame designated ORF 04 was identified between nucleotides 12916-13737 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 28822-29631 of the insert from fosmid 60A5 (SEQ ID NO: 2) on the strands complementary to the insert strands provided in SEQ ID NOs: 1 and 2. These open reading frames have been assigned SEQ ID NOs: 55 and 23 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 56 and 24 respectively in the accompanying sequence listing.

An open reading frame designated Hypothetical 03 was identified between nucleotides 20554-20955 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 37002-37403 of the insert from fosmid 60A5 (SEQ ID NO: 2). These open reading frames have been assigned SEQ ID NOs: 69 and 35 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 70 and 36 respectively in the accompanying sequence listing.

An open reading frame designated ORF 05 was identified between nucleotides 25151-26377 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 39454-40572 of the insert from fosmid 60A5 (SEQ ID NO: 2). These open reading frames have been assigned SEQ ID NOs: 73 and 39 respectively in the accompanying sequence listing, while the polypeptides they encode have been assigned SEQ ID NOs: 74 and 40 respectively in the accompanying sequence listing.

An open reading frame encoding a protein with no homology to known proteins was identified between nucleotides 3-10421 of the insert from fosmid 60A5 (SEQ ID NO: 2). This open reading frame has been assigned SEQ ID No: 3 in the accompanying sequence listing, while the polypeptide it encodes has been assigned SEQ ID No: 4 in the accompanying sequence listing.

An open reading frame designated ORF06 was identified between nucleotides 27535-28002 of the insert from fosmid 101G10 (SEQ ID NO: 1) . This open reading

frame has been assigned SEQ ID No: 77 in the accompanying sequence listing, while the polypeptide it encodes has been assigned SEQ ID No: 78 in the accompanying sequence listing.

A gene coding for $tRNA^{Tyr}$ was identified between nucleotides 12129-12251 of the insert from fosmid 101G10 (SEQ ID NO: 1) and between nucleotides 28058-28180 of the insert from fosmid 60A5 (SEQ ID NO:2) . This tRNA contains a 45 bp intron in the vicinity of the anticodon loop.

Table 1 shows the level of homology between the open reading frames in the inserts from fosmid 101G10 and fosmid 60A5 at the nucleic acid level. Table 1 also shows the level of homology at the amino acid level between the polypeptides encoded by the insert from fosmid 101G10 and fosmid 60A5. Nucleic acid homology was calculated using BLASTN with the default parameters. Amino acid homology was calculated using FASTA with the parameters. As shown in Table 1 and Fig. 1, the protein coding regions were highly similar in both nucleic acid and deduced amino acid sequences.

Over the 28 kb common region in the 101G10 and 60A5 inserts, the inserts shared >99.2% identity in their ribosomal RNA genes, approximately 87.8% overall DNA identity, an average of 91.6% similarity in ORF amino acid sequence, and complete colinearity of protein encoding regions. As shown in Table 1, in protein coding regions the DNA identity of the two contigs ranged from 80.9% (triose phosphate isomerase) to 91.5% (Hypothetical 03). Within intergenic regions the identity dropped to 70 - 86 %, and small insertions or deletions were found frequently. The high similarity in coding regions and upstream sequences aided in the identification of genes, start codons, and putative transcriptional promoter motifs (see below). Genes appear as densely packed in *C. symbiosum* as they are in other sequenced archaeal genomes (Bult, C., *et al.* 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273**, 1058-1073, Klenk, H.P. *et al.* 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature* **390**, 364-370; Smith, D. R., *et al.* 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135-7155).

The ribosomal RNA operon of *Cenarchaeum symbiosum* is composed of the genes for the 16S and 23S rRNAs separated by a spacer of 131 bp. This organization is typical of crenarchaeotes, and differs from rRNA operons of euryarchaeotes, which usually contain 5S RNA and tRNA genes. (Garrett, R. A. *et al.* 1991. Archaeal rRNA operons. *TIBS* **16**, 22-26). The large subunit rRNA genes are located between nucleotides 2680-5674 of SEQ ID NO: 1 (fosmid 101G10) and between nucleotides 18645-21639 of SEQ ID NO: 2 (fosmid 60A5). The small subunit rRNA genes are located between nucleotides 5806-7278 of SEQ ID NO: 1 (on the opposite strand from that shown in the Sequence Listing, as indicated in Figure 1) and between nucleotides 21771-23243 of SEQ ID NO: 2. The large and small subunit rRNA genes in the two fosmids were 99.2% and 99.3% identical, respectively.

As mentioned above, the sequences of the *Cenarchaeum symbiosum* derived inserts in fosmids 101G10 and 60A5 had a high degree of homology but were not completely identical. The sequence of the insert in fosmid 101G10 was designated variant A, while the sequence of the insert in fosmid 60A5 was designated variant B. Such sequence differences could arise if the fosmid inserts were derived from two closely related but distinct strains of *Cenarchaeum symbiosum* or, alternatively, the sequence differences could be due to cloning or sequencing artifacts. To confirm that the fosmid inserts were in fact derived from two closely related strains, portions of the inserts in a plurality of different fosmids were sequenced to determine whether they were identical to either of the inserts in fosmids 101G10 and 60A5, as would be the case if there were in fact two closely related strains of *Cenarchaeum symbiosum.*

In particular, the ribosomal RNA spacer regions of variant A and variant B contained 10 distinguishing signature nucleotides and the 16S rRNA genes of variant A and variant B contained two distinguishing nucleotides. Example 5 provides the results of a PCR based analysis of the 16S rRNA gene and the 16S-23S spacer region in 13 different fosmid inserts.

Example 5

PCR Based Analysis of Fosmid Inserts to Determine Whether they Contain the Variant A or Variant B Sequences

Primers 21F and 459R-LSU (CTTTCCCTCACGGTA, SEQ ID NO: 116) were used to amplify the 16S-23S spacer region from the fosmids. The amplification

products were sequenced using primer SP23rev (CTA TTG CCG TCT TTA CACC, SEQ ID NO: 117).

PCR reactions with two archaea-specific 16S rDNA primers (21F and 958R (DeLong, E. F. 1992. Archaea in coastal marine environments. *Proc. Natl. Acad. Sci.* **89**, 5685-5689, the disclosure of which is incorporated herein by reference), one of which was biotinylated, were used to amplify a 950 base pair (bp) fragment from the fosmids. The PCR products were purified and sequenced as described in Preston, C. M. *et al.* 1996. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc. Natl. Acad. Sci.* USA **93,** 6241-6246 with primer 519R 16S rDNA

The results of this analysis are shown in Table 2. As shown in Table 2, in samples obtained from several unique rRNA operon-containing fosmids, a sequence identical to either variant A (101G10) or variant B (60A5) was present.

The above methods may also be used to determine whether a biological sample contains variant A and/or variant B. In such procedures, nucleic acids are obtained from the biological sample, amplified using the above primers, and sequenced using the above oligonucleotide to determine whether the sample contains the variant A and/or the variant B sequence.

Similarly, the amplification reaction may be conducted using any primers which generate amplification products having sequences which differ between variant A and variant B. The amplification products may then be sequenced to determine whether they have the sequence of variant A and/or variant B. In some embodiment, the amplification reaction may be conducted under conditions in which the amplification primers specifically hybridize to one of the variants.

RFLP analyses were also be used to assess whether the fosmids contained the sequence of variant A or variant B as described in Example 6 below.

Example 6

RFLP Based Analysis of Fosmids to Determine Whether They Contain the Variant A or Variant B Sequences

Primer set 21F (DeLong, E. F. 1992. Archaea in coastal marine environments. *Proc. Natl. Acad. Sci.* **89**, 5685-5689) and 459R-LSU for the amplification of 2.2 kbp of

the ribosomal operon, primer set GSAT810F (GAATCCGCC CCCGACTATCTT, SEQ ID NO: 118) and 16S37REV (CATGGCTTAGTATCAATC SEQ ID NO: 119) for the amplification of the 16S RNA-GSAT region (2.2 kbp) and primer set Cenpol357F (ACITACAACGGI GACGAYTTTGA SEQ ID NO: 120) and Cenpol735R (CACCCCGAARTAGTTYTTYTT SEQ ID NO: 121) for an internal DNA polymerase fragment (of 1134 bp) were used in PCR reactions with 5 ng of purified fosmids. The PCR products were cut with TaqI and HpaII (16S-23S RNA), HaeIII and RsaI (GSAT-16S RNA) or HaeIII and AvaII (polymerase) and analyzed on 2 % agarose gels.

The results are shown in Table 2. If the pattern did not exactly match but closely resembled the RFLP of either type A or B, it was assigned as a lower case letter (a or b, Table 2), meaning that at least 3 out of 4 or 3 out of 5 bands created by restriction digest appear identical in size to the ones from either type A or B. As shown in Table 2, RFLP patterns of the 1150 bp fragment covering the 5'-end of the GSAT gene and 16S gene and the internal fragment of 1134 bp from the DNA polymerase gene revealed that all fosmids analyzed could again be assigned to either the A or B type, although slight variations were also detected (lower case letters in Table 2), suggesting that both variants exhibit further microheterogeneity which is detectable in protein coding and intergenic regions.

The above methods may also be used to determine whether a biological sample contains variant A and/or variant B. In such procedures, nucleic acids are obtained from the biological sample, amplified using the above primers, and digested as described above to determine whether the sample contains the variant A and/or the variant B sequence. Similar analyses may also be performed using other portions of the sequences of SEQ ID NOs: 1 and 2 which are different from one another.

To further confirm the existence of two closely related strains of *Cenarchaeum symbiosum*, biological samples were obtained from several individual sponges and analyzed to determine whether the samples contained variant A and/or variant B. Example 7 below provides the results of a PCR analysis of the *Cenarchaeum symbiosum* 16S rRNA genes in samples obtained from several individual sponges in different locations and at different times.

Example 7

Analysis of Samples from Individual Sponges

The 16S rRNA genes of variant A and variant B differ at positions 175 and 183.7 (*E. coli* numbering). PCR reactions with two archaea-specific 16S rDNA primers (21F and 958R (DeLong, E. F. 1992. Archaea in coastal marine environments. *Proc. Natl. Acad. Sci.* **89**, 5685-5689, the disclosure of which is incorporated herein by reference), one of which was biotinylated, were used to amplify a 950 base pair (bp) fragment from total nucleic acids derived from several different sponge individuals. The PCR products were purified and sequenced as described in Preston, C. M. *et al.* 1996. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc. Natl. Acad. Sci.* USA **93,** 6241-6246 with primer 519R, the disclosure of which is incorporated herein by reference.

The amplification products were sequenced to determine whether they corresponded to variant A and/or variant B. The results are shown in Table 3. As shown in Table 3, in 15 out of 16 cases U/C ambiguities were found at the signature positions, indicating the presence of both variants in samples obtained from a single sponge (Table 3). Only one sponge (S4) yielded an unambiguous sequence identical to variant A, but variant B was detected in this individual by another criterion (see below).

Hybridization analyses were also used to determine whether individual sponges harbored variant A and/or variant B. The results of these analyses are provided in Example 8 below.

Example 8

Hybridization Based Analysis of Samples Obtained from *Axinella Mexicana* to Determine Whether the Samples Contain Variant A and/or Variant B

Two oligonucleotides specific for each variant type were designed from the 23S rDNA gene sequences of fosmids 101G10 and 60A5. The probes differed in 3 positions and have the sequences ACACTTCAACTATTTCCTG (SEQ ID NO: 122 variant A) and ACACTTTGACTATTTCGTG (SEQ ID NO: 123, variant B). Nucleic acid samples from individual sponges (300 ng) and controls (fosmids 101G10 and 60A5, 50 ng each) were denatured, bound to nylon membranes (Hybond-N, Amersham), hybridized with the labeled probes (Massana, R. *et al.* 1997. Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara

Channel. *Appl. Env. Microb.* **63**, 50-56, the disclosure of which is incorporated herein by reference in its entirety) and washed at 41.5 °C. Hybridization was analyzed by autoradiography.

The results are provided in Table 3. In the samples from the majority of host sponges examined, the presence of both 23S rRNA variants was observed, confirming that the specific association of *C. symbiosum* with its host typically involves the presence of both variants.

The data provide strong evidence that these genomic clones are derived from two very closely related, but distinct strains, as opposed to representing two ribosomal RNA operon regions originating from the same organism. This conclusion is consistent with the observation that all crenarchaeota characterized to date contain only one ribosomal RNA operon (Garrett, R. A. *et al.* 1991. Archaeal rRNA operons. *TIBS* **16**, 22-26).

The high conservation between the inserts in fosmid 101G10 and fosmid 60A5 was not entirely confined to coding regions but also extended into adjacent upstream sequences. Due to this upstream similarity, and also because the average G+C content of the sequences was relatively high, it was possible to readily identify prospective transcriptional (A+T rich) promoter elements. A motif corresponding to the consensus of the archaeal TATA-box-like element (C/T-T-T-A-T/A-A) (Hain, J. *et al.* 1992. Elements of an archaeal promoter defined by mutational analysis. *Nucl. Acids. Res.* **20**, 5423-5428) was identified upstream of nearly all genes (Fig. 2). The exceptions were the genes encoding MenA and DNA polymerase which are located immediately downstream of other ORFs and may therefore be transcribed as polycistronic mRNAs. *In vivo* and *in vitro* studies in other archaea have shown that initiation of transcription occurs consistently 24 to 28 bp downstream from the central T of this motif (Hain, J *et al.* 1992. Elements of an archaeal promoter defined by mutational analysis. *Nucl. Acids. Res.* **20**, 5423-5428; Palmer, J. R. and Daniels, C.J. 1995. In vivo definition of an archaeal promoter. *J. Bacteriol.* **177** 1844-1849). For twelve of the protein encoding genes, the promoter element was found 25 to 30 bp upstream of the ORF (Fig. 2), suggesting that transcriptional initiation occurs in close proximity to, or directly at, the translational start codon.

A similar observation has been made for 30 of the predicted 100 strong and medium promoters from 156 kbp sequence of *Sulfolobus solfataricus* (Sensen, C. W. *et al.* 1996. Organizational characteristics and information content of an archaeal genome: 156 kb of sequence from *Sulfolobus solfataricus* P2. *Molec. Microb.* **22**, 175-191). Transcription initiation at, or in close proximity to, the translational start codons has been mapped for some genes in *Halobacterium salinarium* (Brown, J.W. *et al.* 1989. Gene structure, organization, and expression in archaebacteria. *CRC Crit. Rev. Microb.* **16**, 287-337) and *S. solfataricus* (Klenk, H.P., *et al.* 1993. Nucleotide sequence, transcription and phylogeny of the gene encoding the superoxide dismutase of Sulfolobus acidocaldarius. *Biochim. Biophys. Acta* **1174** 95-98), and alternative mechanisms for initial mRNA-ribosome contact in *Archaea* have been hypothesized (Brown, J.W. *et al.* 1989. Gene structure, organization, and expression in archaebacteria. *CRC Crit. Rev. Microb.* **16**, 287-337).

The promoters listed in Figure 2, or fragments thereof, may be used in expression vectors or expression systems. In one embodiment, the promoters listed in Figure 2 may be operably linked to coding regions and introduced into archaebacteria, and in particular *Cenarchaeum symbiosum*, to express the encoded gene product in the archaebacterial cells.

Alternatively, the promoters listed in Figure 2 may be operably linked to coding regions and introduced into host cells which are not normally capable of directing transcription from archaebacterial promoters. In addition, genes encoding the proteins required for transcription from these promoters are also introduced into the host cells. The genes encoding these transcription factors may be on the same vector as the promoter from *Cenarchaeum symbiosum* or on a different vector. In some embodiments, the genes encoding these transcription factors are linked to an inducible promoter. Expression of the transcription factors is induced when it is desired to express the proteins which are operably linked to the promoter from *Cenarchaeum symbiosum*.

Although this invention has been described in terms of certain preferred embodiments, other embodiments which will be apparent to those of ordinary skill in the art in view of the disclosure herein are also within the scope of this invention.

Accordingly, the scope of the invention is intended to be defined only by reference to the appended claims. All documents cited herein are incorporated herein by reference in their entirety.

Table 1

Comparison of Overlapping Coding Sequences from Fosmid 101G10 and Fosmid 60A5

Gene Name[1]	Functional Category	% Identity	
		Nucleotide	Amino Acid
Hypothetical 01	unknown	81.4	76.6
23S	translation	99.16	
16S	translation	99.3	
GSAT	heme biosynthesis	83.2	83.8
Hypothetical 02	unknown	83.4	81.4
ORF 01	unknown	83.3	85.7
ORF 02	unknown	89.9	95.2
ORF 03	unknown	87.9	86.7
$tRNA^{tyr}$	translation	99.2	
ORF 04	unknown	87.8	88.1
TIM	glycolysis	80.9	83.3
TBP	transcription	83.4	86.3
DNA polymerase	replication/repair	89.0	93.9
dCMP deaminase	pyrimidine synthesis	85.7	89.8
RNA helicase (ATP dependent)	translation	86.1	92.2
PPI	chaperone	88.4	92.5
Hypothetical 03	unknown	91.5	92.4
MenA	menaquinone biosynthesis	86	89.4
ORF 05	unknown	87.5	90.6
Methylase	restriction/modification	86.4	87.5

[1] Hypothetical: open reading frame (ORF) with similarity to proteins of unknown function from the databases.

ORF = open reading frame identified by similarity between both fosmids, including upstream promoter sequence; GSAT = glutamate semialdehyde aminotransferase; TIM = triose-phosphate isomerase; TBP = TATA box-binding protein; PPI = peptidylprolyl cis/trans isomerase.

Table 2

Analysis of Polymorphism at Four Distinct Loci in Different Fosmids

Fosmid	16S RNA*1	16S-23S spacer*2	16S-GSAT*3		DNA Pol*3	
			HaeIII	RsaI	HaeIII	AvaII
101G10	A	A	A	A	A	A
60A5	B	B	B	B	B	B
15A5	B	B	--	--	b	b
43H4	A	--	--	--	A	A
60H6	A	A	--	--	a/b	B
69H2	A	--	--	--	A	A
87F4	B	--	--	--	b	a/b
C1H5	A	A	A	A		
C4H1	A	A	A	A		
C4H9	A	A	A	A	A	B
C7D4	A	A	A	A	A	A
C8B8	B	B	B	B	B	b
C15A3	A	A	A	A		
C17D2	B	--	b	B	B	b
C20B5	A	A	a	a/b		

*1: partial sequence (101G10 through 87F4) or RFLP analysis (C1H5 through C20B5).
*2: partial sequence.
*3: RFLP analysis of PCR products; A/B: identical pattern to either 101G10 (=A) or 60A5 (=B); a,b: similar pattern to either A or B (see materials and methods). Fosmids C1H5, C4H1, C15A3 and C20B5 did not yield PCR products with polymerase-specific primers.
The first seven fosmids were isolated from a first library, the last 8 fosmids (prefix C) are from a second library.
-- = not determined.

Table 3

Detection of *C. symbiosium* Variants in Natural Populations of *A. mexicana*

A. mexicana Individual or Isolated DNA Source*	Variation in 16S rDNA Positions**		Variations in 23S rRNA Hybridization	
	175	183.7	Variant Type A	Variant Type B
fosmid 101G10 from s12	U	U	+	-
fosmid 60A5 from s12	C	C	-	+
s12	Y	Y	+	+
s1	---	---	+	+
s2	---	---	+	+
s3	Y	Y	+	+
s4	U	U	+	w
s5	Y	Y	---	---
s6	Y	Y	+	+
s7	---	---	+	w
s8	Y	Y	+	+
s9	Y	Y	+	w
s10	---	---	+	+
s11	Y	Y	+	+
s13	---	---	+	+
s14	---	---	+	w
s16	---	---	+	+
s17	---	---	-	w
s18	Y	Y	-	w
s19	---	---	+	+
s20	---	---	+	+
s21	---	---	+	+
s22	---	---	+	+
s23	---	---	+	+
s24	---	---	+	+
s25	---	---	+	+
s26	---	---	+	+
s27	---	---	+	+
s28	---	---	+	+
s29	---	---	+	+
s30	---	---	+	+
hs1	---	---	+	+
hs2	---	---	+	+
hs3	Y	Y	+	w
hs4	Y	Y	+	w
hs5	Y	Y	+	+
hh1	---	---	w	w
hh2	Y	Y	+	+
hh3	Y	Y	+	+
Aq1	Y	Y	---	---
Aq2	Y	Y	---	---
Aq3	---	---	+	+

*s = Naples Reef; hs = Haskle; hh = Hermit Hole; Aq = captive sponge.
**Y = direct sequence of PCR product yields C and U at the same position.
--- = not determined; w = weakly positive.